TÉCNICO+
FORMAÇÃO AVANÇADA

Deep Learning

Gonçalo M. Correia

GALP 2023

Program for today

Generative models

Trust, interpretability, ethics

So far...

- All models discussed so far addressed supervised learning

So far...

- Example: Feed-forward neural networks
  - Data used:

$$D = \{(x_n, y_n), n = 1, \ldots, N\}$$

    where x_n is some vector of features, and y_n is the desired output (label, value)

  - Network uses data to learn a correspondence between input space, $\mathcal{X}$, and output space, $\mathcal{Y}$

So far...

- Example: Convolutional neural networks
  - Data used:

$$D = \{(x_n, y_n), n = 1, \ldots, N\}$$

    where x_n is some structured input (e.g., image, sound), and y_n is the desired output (label, value)

  - Network uses data to learn a correspondence between $\mathcal{X}$ and $\mathcal{Y}$

So far...

- Example: Recurrent neural networks
  - Data used:

$$D = \{(x_n, y_n), n = 1, \ldots, N\}$$

    where x_n is a sequence of inputs (e.g., sentence), and y_n is the desired output (e.g., next word)

  - Network uses data to learn a correspondence between $\mathcal{X}$ and $\mathcal{Y}$

So far...

- Example: Sequence-to-sequence models
  - Data used:

$$D = \{(x_n, y_n), n = 1, \ldots, N\}$$

    where x_n is a sequence of inputs (e.g., sentence in a language), and y_n is a desired output sequence (e.g., sequence in a different language)

  - Network uses data to learn a correspondence between $\mathcal{X}$ and $\mathcal{Y}$

So far...

- Some of these models can be trained in a self-supervised manner
- The same data is used for input and for output
- Examples:
  - Auto-encoders
  - (Static) word embeddings
  - Large pre-trained models

So far...

- Some of these models can be trained in a self-supervised manner
- The same data is used for input and for output
- Self-supervised learning can be seen as a very particular form of unsupervised learning

Generative models

- What if we want to model the data itself?
- In other words, given a dataset

$$D = \{x_n, n = 1, \ldots, N\}$$

we want to learn a model for $x_n, n = 1, \ldots, N$.

Generative models

- A generative model is exactly a model of the data itself
- It is called "generative" since we can use it to generate synthetic data
- The model typically consists of learning a function

$$p_w : \mathcal{X} \to [0, 1]$$

so that

$$p_w(x) \approx \mathbb{P}[X = x]$$

Examples of generative models

- Auto-regressive networks
- Restricted Boltzmann machines
- Deep belief networks
- Deep Boltzmann machines
- Variational auto-encoders (VAEs)
- Generative adversarial networks (GANs)
- ...

Examples of generative models

- Auto-regressive networks
- Restricted Boltzmann machines
- Deep belief networks
- Deep Boltzmann machines
- Variational auto-encoders (VAEs)
- Generative adversarial networks (GANs)
- ...

Differentiable generator networks

- Both VAEs and GANs belong to a class of models known as differentiable generator networks
- These networks compute a conditional model $p_w(x \mid z)$
  - z is a (low-dimensional) latent variable
  - z captures different dimensions of variation of the data

How can we train p_w?

Let's play the math a bit...

- From Bayes rule,

$$p_w(z \mid x) = \frac{p_w(x, z)}{p_w(x)}$$

or, equivalently,

$$p_w(x) = \frac{p_w(x, z)}{p_w(z \mid x)}$$

Let's play the math a bit...

- From Bayes rule,

$$p_w(z \mid x) = \frac{p_w(x, z)}{p_w(x)}$$

or, equivalently,

$$\log p_w(x) = \log p_w(x, z) - \log p_w(z \mid x)$$

Let's play the math a bit...

- Let's take an arbitrary distribution q over z...

$$\log p_w(x) = \log p_w(x, z) - \log p_w(z \mid x)$$

Let's play the math a bit...

- Let's take an arbitrary distribution q over z...

$$\log p_w(x) = \log p_w(x, z) - \log p_w(z \mid x) + \log q(z) - \log q(z)$$

Let's play the math a bit...

- Let's take an arbitrary distribution q over z...

$$\log p_w(x) = \log p_w(x, z) - \log q(z) + \log \frac{q(z)}{p_w(z \mid x)}$$

- Taking the expectation w.r.t. q...

$$\boxed{\mathbb{E}_{z\sim q}[\log p_w(x)]} = \mathbb{E}_{z\sim q}[\log p_w(x, z) - \log q(z)] + \boxed{\mathbb{E}_{z\sim q}\left[\log \frac{q(z)}{p_w(z \mid x)}\right]}$$

Does not depend on z

Non-negative

Let's play the math a bit...

- Let's take an arbitrary distribution q over z...

$$\log p_w(x) = \log p_w(x, z) - \log q(z) + \log \frac{q(z)}{p_w(z \mid x)}$$

- Taking the expectation w.r.t. q...

$$\log p_w(x) \geq \boxed{\mathbb{E}_{z \sim q}[\log p_w(x, z) - \log q(z)]}$$

Evidence lower bound (ELBO)

ELBO

- Evidence lower bound (ELBO):

$$\log p_w(x) \geq \mathbb{E}_{z \sim q}[\log p_w(x, z) - \log q(z)]$$

- Idea:
  - If we push the right-hand side up, we push the left-hand side up
- How?
  - Choosing q adequately

ELBO

- We can rewrite the ELBO as

$$\text{ELBO} = \mathbb{E}_{z \sim q}[\log p_w(x \mid z)] - \text{KL}(q(z) \parallel p_{\text{prior}}(z))$$

Maximize the ability to recover x from z (reconstruction)

Minimize the distance between q and prior (regularization)

What about q?

- How can we select q?
- Standard approach is to pick q to...
  - ... select z from x
  - ... be some "friendly distribution" (e.g., Gaussian)
  - ... be itself a neural network

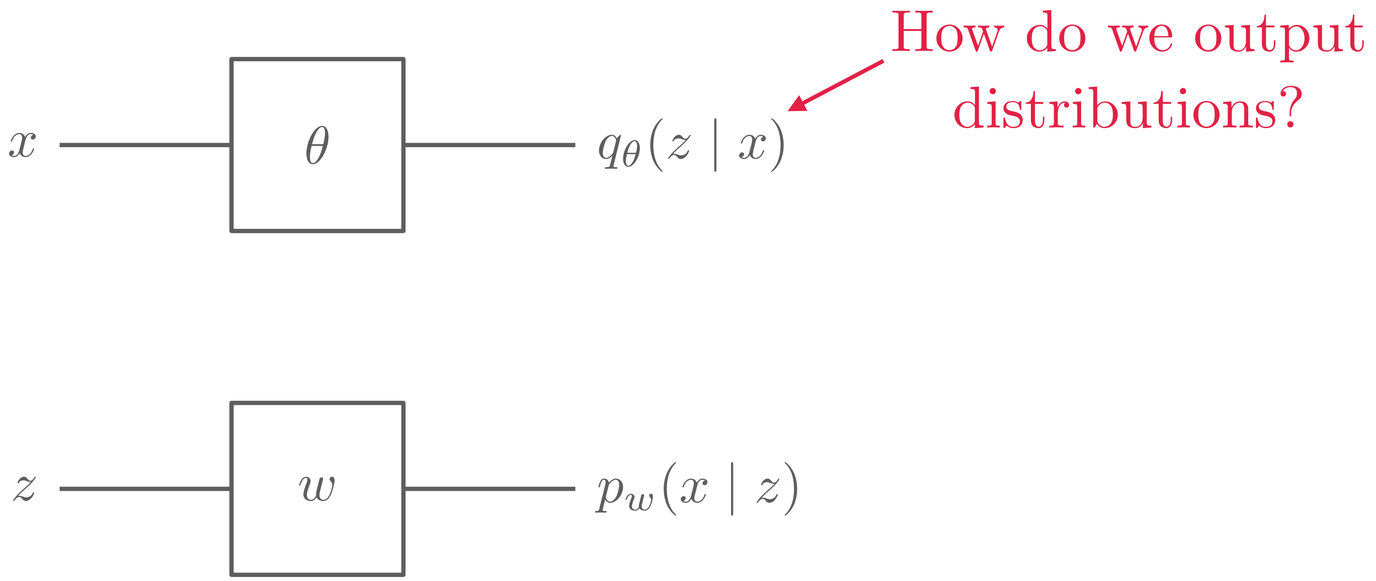
Putting everything together...

- Back to the ELBO:

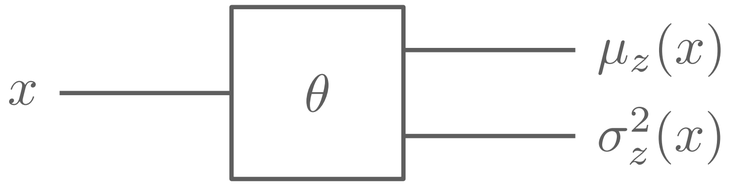
$$\text{ELBO} = \mathbb{E}_{z \sim q}[\log p_w(x \mid z)] - \text{KL}(q_\theta(z \mid x) \,\|\, p_{\text{prior}}(z))$$

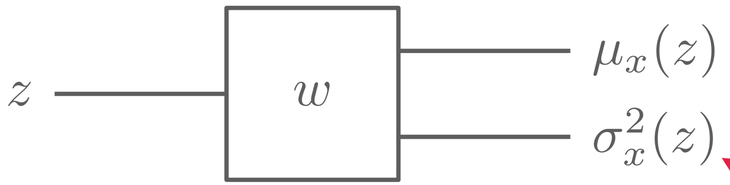
- We have 3 elements:
  - A neural network that encodes $p_w(x \mid z)$
  - A neural network that encodes $q_\theta(z \mid x)$
  - The prior p_{prior}, usually a Normal$(0, I)$

Putting everything together...



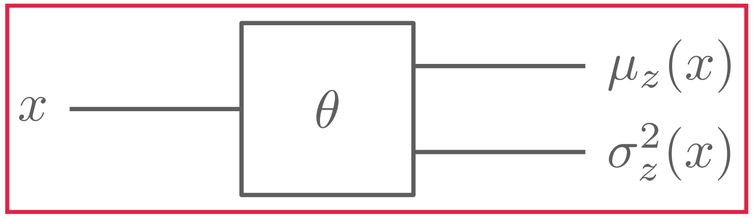
Putting everything together...



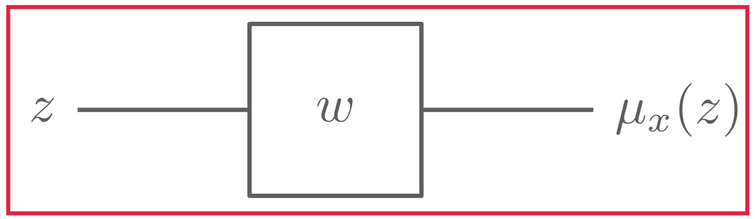


To make our life easier,
we usually consider
$\sigma_x^2(z) = 1$

Putting everything together...

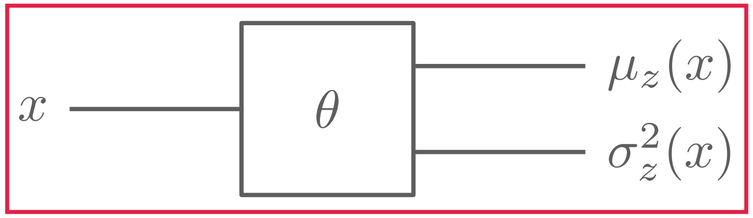


Returns a (sampled) representation z of x

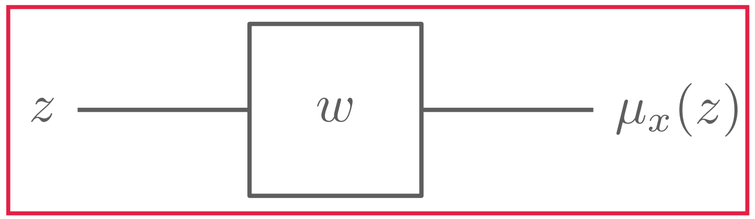


Recovers x from z

Putting everything together...

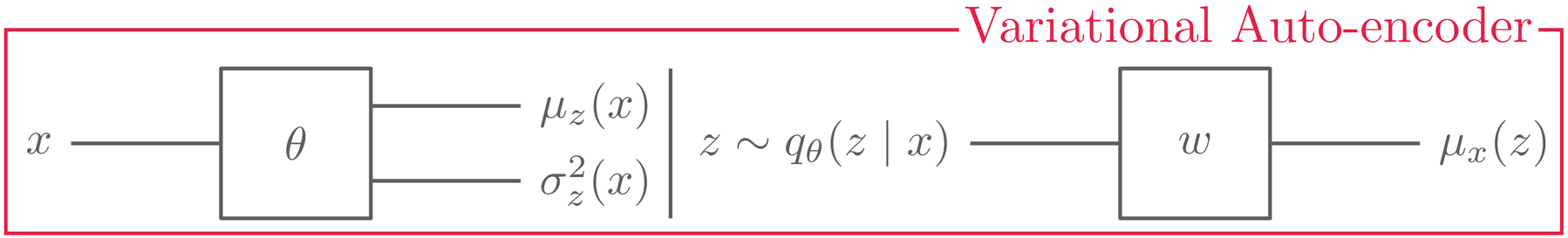


Encoder

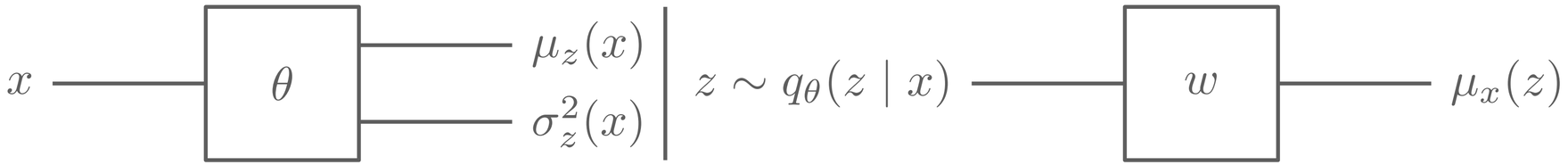


Decoder

Putting everything together...



Training a VAE



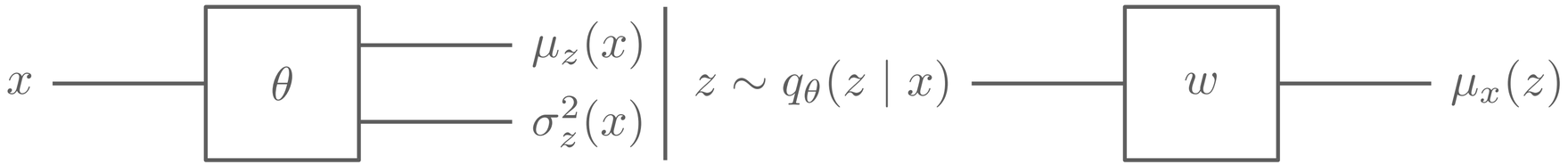
- We want to maximize the ELBO:

$$\text{ELBO} = \mathbb{E}_{z\sim q}[\log p_w(x \mid z)] - \text{KL}(q_\theta(z \mid x) \parallel p_{\text{prior}}(z))$$

or, equivalently, minimize the loss

$$L(w, \theta) = \text{KL}(q_\theta(z \mid x) \parallel p_{\text{prior}}(z)) - \mathbb{E}_{z\sim q}[\log p_w(x \mid z)]$$

Training a VAE



- We want to maximize the ELBO:

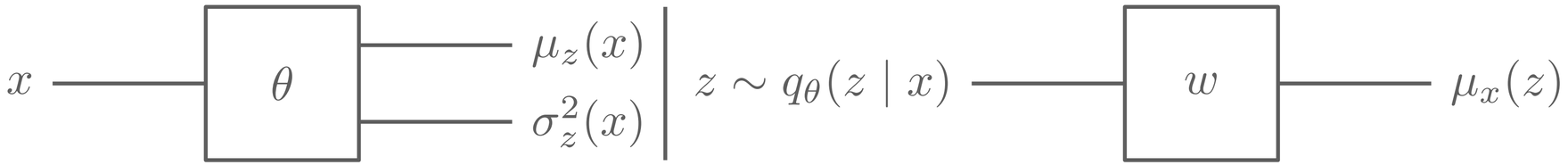
$$\text{ELBO} = \mathbb{E}_{z\sim q}[\log p_w(x \mid z)] - \text{KL}(q_\theta(z \mid x) \parallel p_{\text{prior}}(z))$$

or, equivalently, minimize the loss

$$L(w, \theta) = \mathbb{E}_{z\sim q}[\log q_\theta(z \mid x) - \log p_{\text{prior}}(z) - \log p_w(x \mid z)]$$

How do we deal with the expectation?

Training a VAE



- We want to maximize the ELBO:

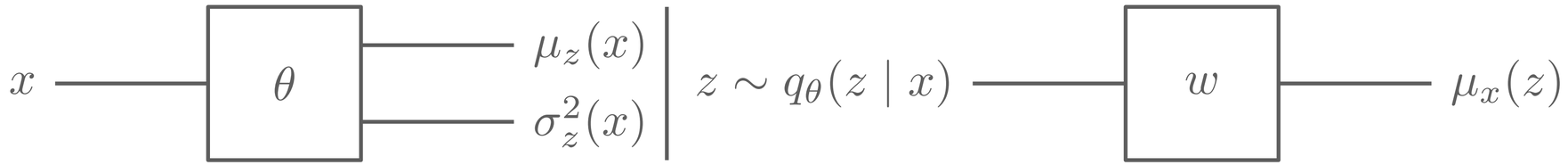
$$\text{ELBO} = \mathbb{E}_{z\sim q}[\log p_w(x \mid z)] - \text{KL}(q_\theta(z \mid x) \parallel p_{\text{prior}}(z))$$

or, equivalently, minimize the loss

$$L(w,\theta) = \frac{1}{M}\sum_{m=1}^{M}(\log q_\theta(z_m \mid x) - \log p_{\text{prior}}(z_m) - \log p_w(x \mid z_m))$$

Usually, $M = 1$

Training a VAE



- We want to maximize the ELBO:

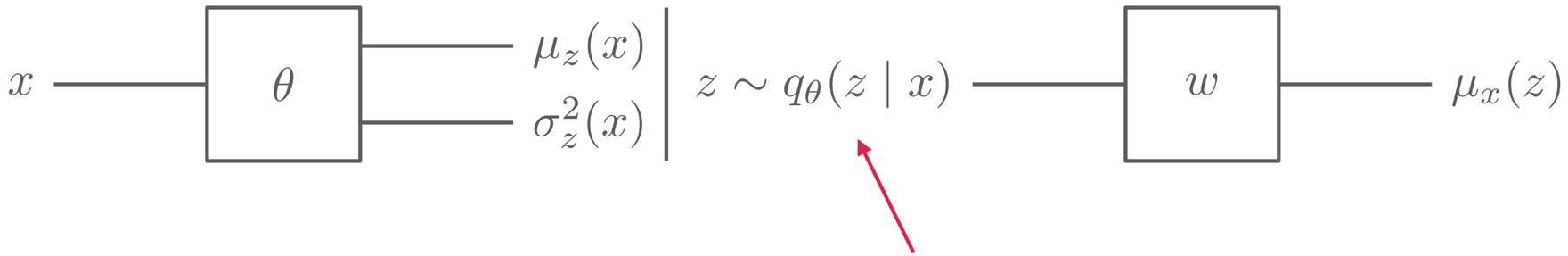
$$\text{ELBO} = \mathbb{E}_{z\sim q}[\log p_w(x \mid z)] - \text{KL}(q_\theta(z \mid x) \parallel p_{\text{prior}}(z))$$

or, equivalently, minimize the loss

$$L(w, \theta) = \log q_\theta(z \mid x) - \log p_{\text{prior}}(z) - \log p_w(x \mid z)$$

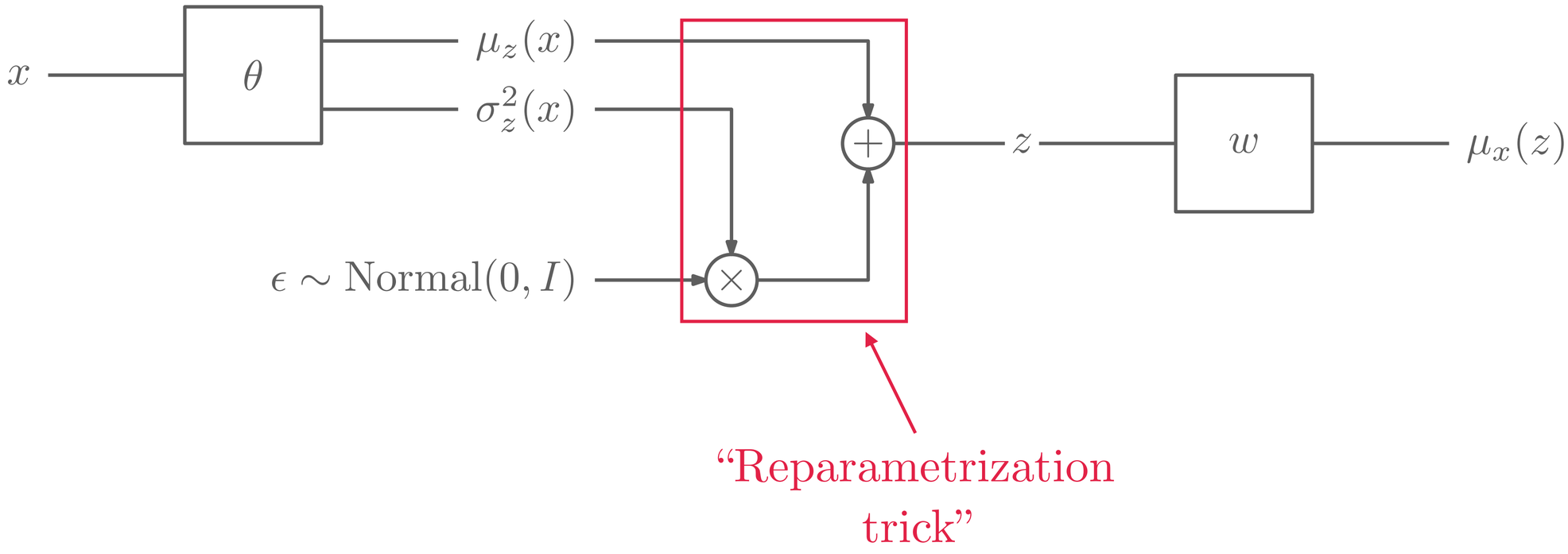
with $z \sim q_\theta(\cdot \mid x)$

Training a VAE



How do we backprop through sampling??

Training a VAE

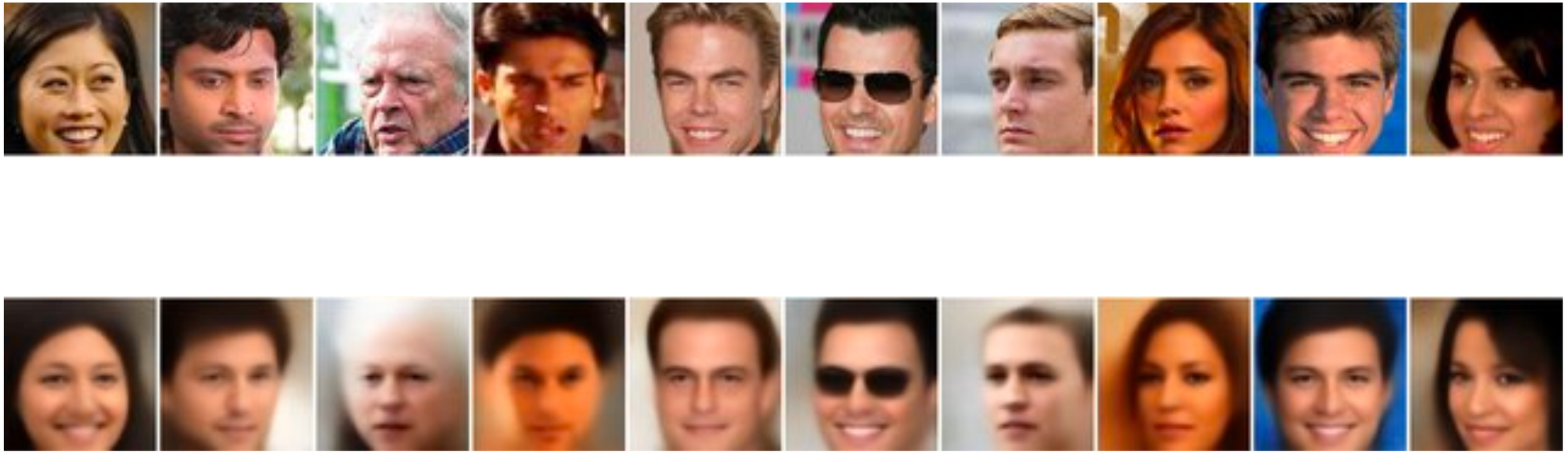


Reconstruction using VAE

Original images

Reconstructed images

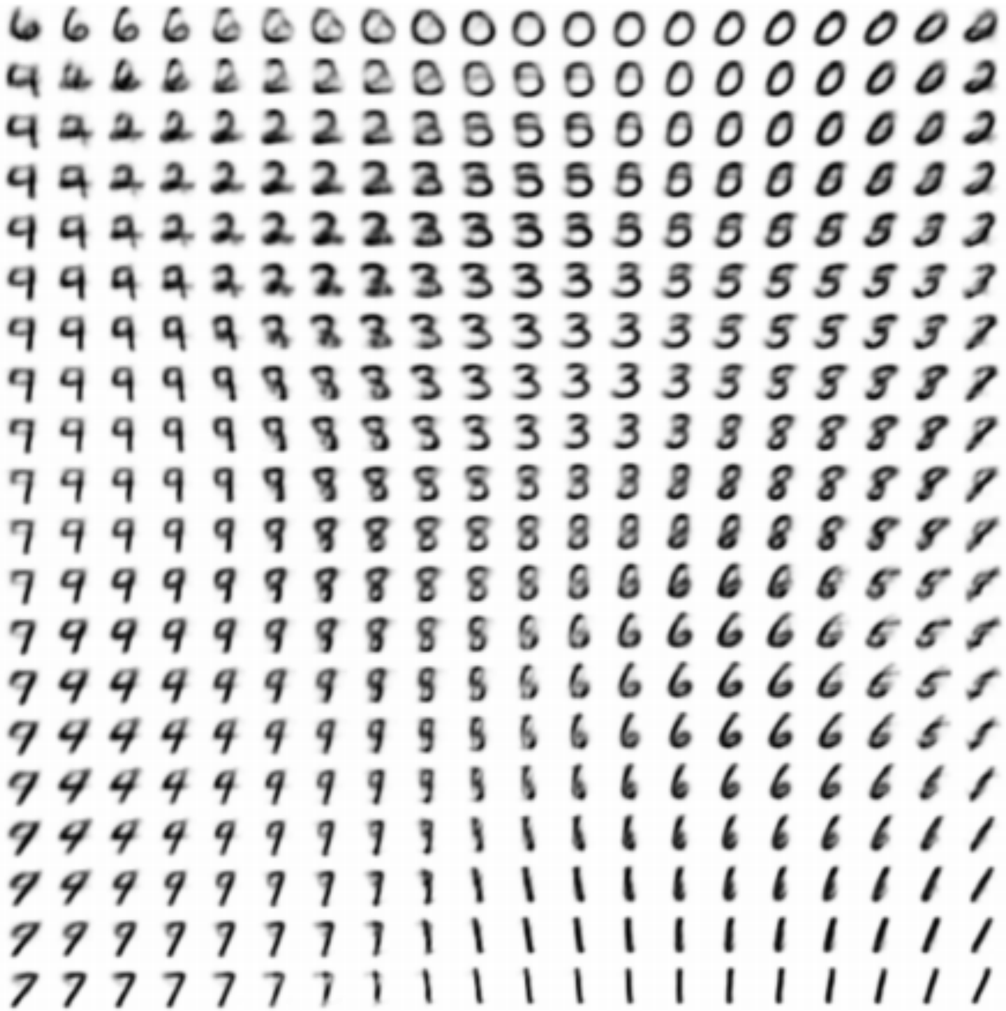
Reconstruction using VAE



Generation using a VAE

- We can now use the trained decoder to generate artificial samples:
  - We sample $z \sim p_{\text{prior}}$
  - We use the decoder to generate synthetic samples $\hat{x}$ from z, $\hat{x} \sim p_w(\cdot \mid z)$

Generation using a VAE



Generated images as latent space is traversed

Generation using a VAE

Generated images as latent space is traversed

How can we train p_w? (reprise)

Training p_w

- VAEs depart from a Bayesian approach to training p_w
  - The approach requires "friendly" distributions (e.g., Gaussians)
  - The resulting images are "blurry"
  - VAEs may suffer from posterior collapse (i.e., learning to ignore latent variable)

Training p_w

- If the goal is to generate realistic data, why not evaluate the data on "realism"?
- Idea:
  - Train our generator network to produce realistic synthetic data
  - Train a second network (discriminator) to distinguish real from synthetic data

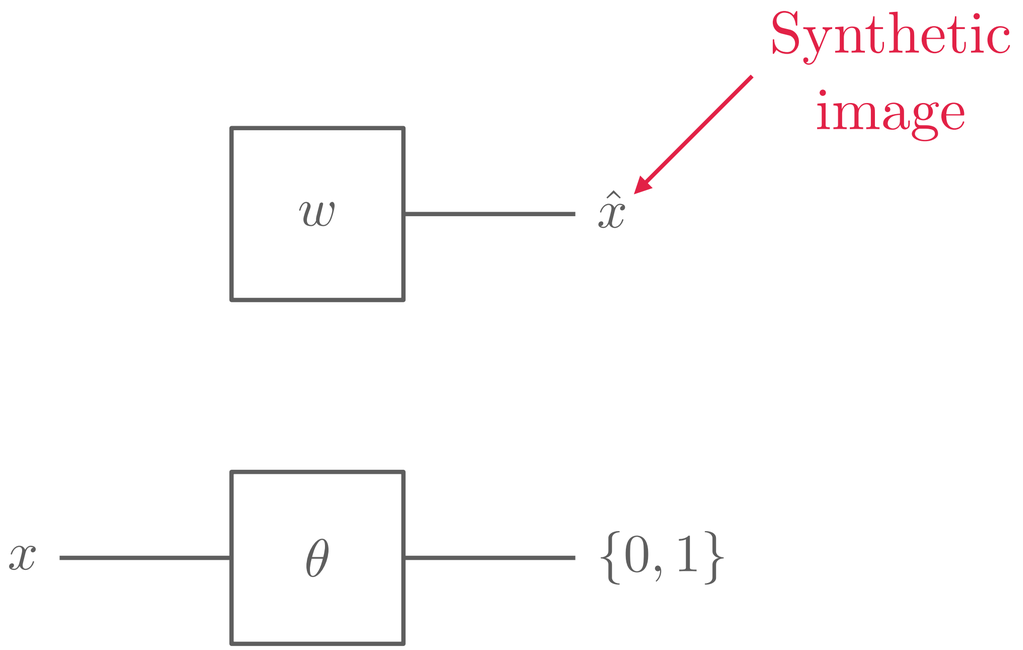
Performance: Ability to fool the discriminator

Performance: Ability to recognize real from fake images

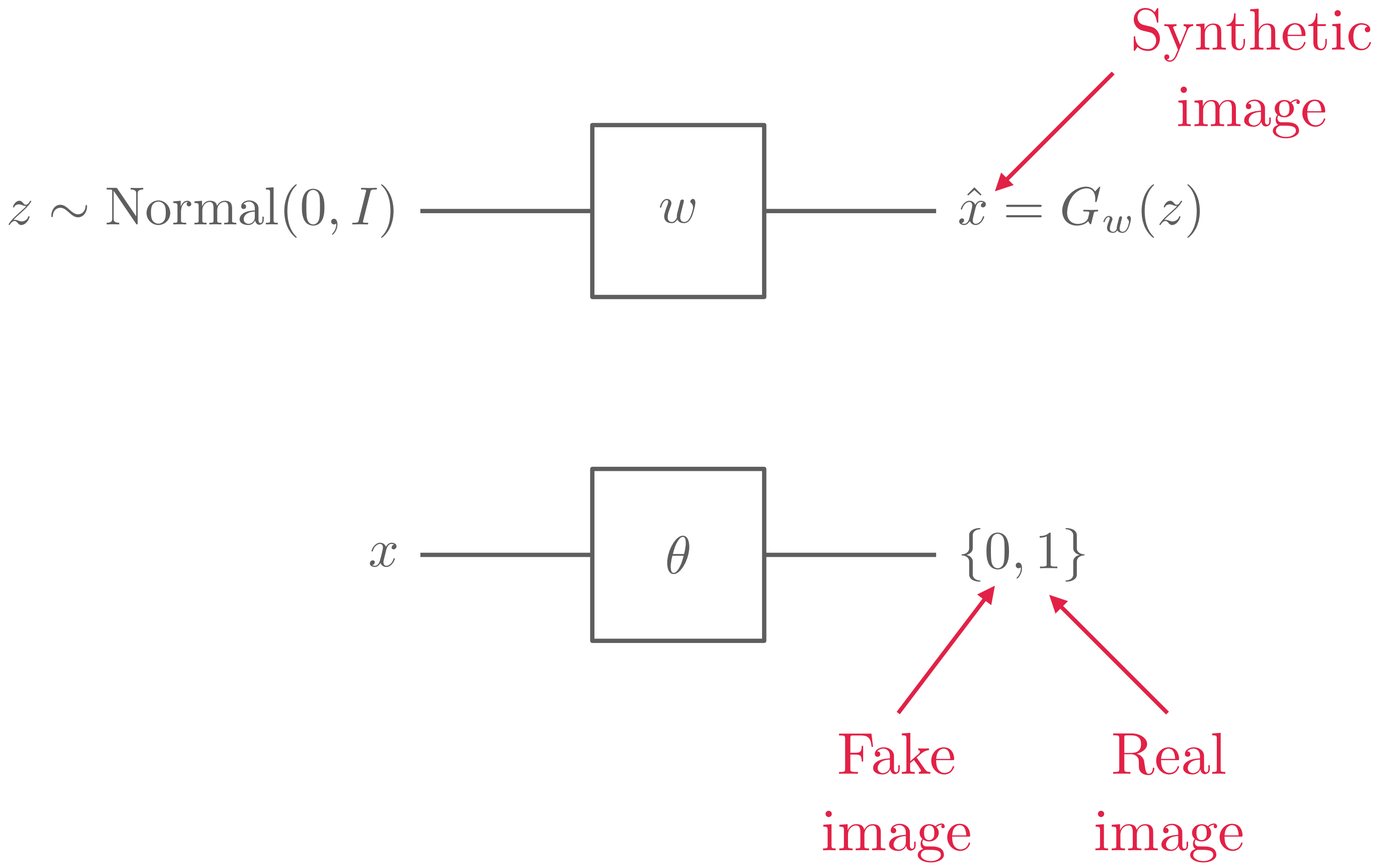
Training p_w

- If the goal is to generate realistic data, why not evaluate the data on "realism"?
- Idea:
  - Train our generator network to produce realistic synthetic data
  - Train a second network (discriminator) to distinguish real from synthetic data
- The training process is adversarial (the two networks are "competing" with one another)

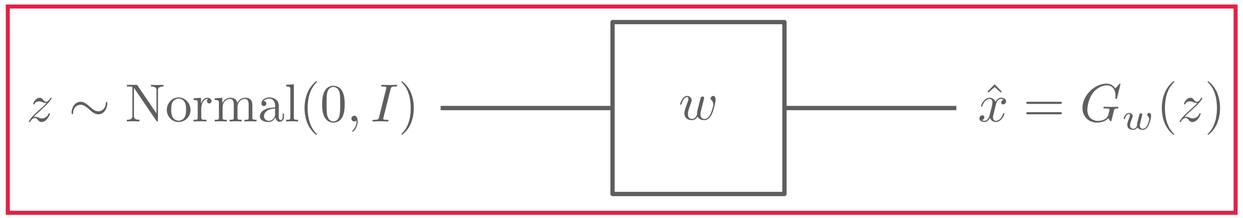
Generative adversarial network (GAN)



Generative adversarial network (GAN)



Generative adversarial network (GAN)



Decoder
(generator)

Classifier
(discriminator)

Generative adversarial network (GAN)

- Generator network, G, and discriminator network, D, are playing a game
  - Discriminator wants to maximize

$$\mathbb{E}_{x \sim \mu_{\text{data}}} [\log p_\theta(x)]$$

Real data

and minimize

$$\mathbb{E}_{x \sim p_w(\cdot|z)} [\log p_\theta(x)]$$

"Fake" data

Generative adversarial network (GAN)

- Generator network, G, and discriminator network, D, are playing a game
  - Discriminator wants to maximize

$$\mathbb{E}_{x \sim \mu_{\text{data}}} \left[\log p_\theta(x) \right]$$

and minimize

$$\mathbb{E}_{z \sim \text{Normal}(0, I)} \left[\log p_\theta(G_w(z)) \right]$$

Generative adversarial network (GAN)

- Generator network, G, and discriminator network, D, are playing a game
  - Discriminator wants to maximize

$$\mathbb{E}_{x \sim \mu_{\text{data}}} \left[\log p_\theta(x) \right]$$

and

$$\mathbb{E}_{z \sim \text{Normal}(0, I)} \left[\log(1 - p_\theta(G_w(z))) \right]$$

Generative adversarial network (GAN)

- Generator network, G, and discriminator network, D, are playing a game
  - Generator wants to minimize

$$\mathbb{E}_{x \sim \mu_{\text{data}}} \left[\log p_\theta(x) \right]$$

and

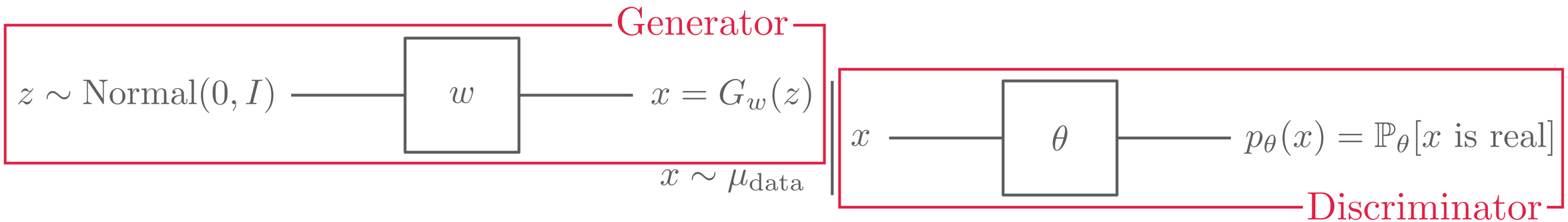
$$\mathbb{E}_{z \sim \text{Normal}(0, I)} \left[\log(1 - p_\theta(G_w(z))) \right]$$

Generative adversarial network (GAN)

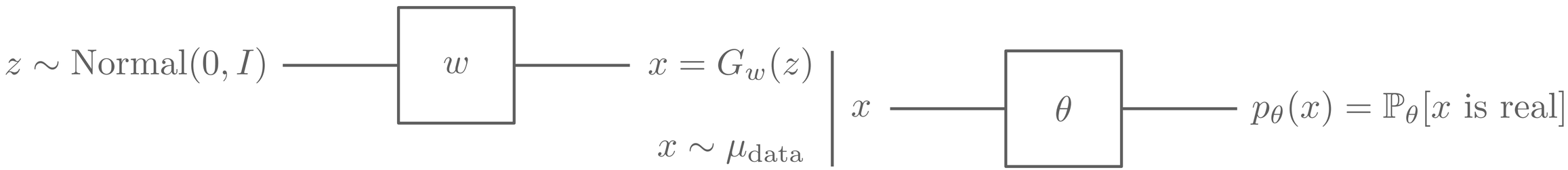
- Generator network, G, and discriminator network, D, are playing a game
- Saddle point problem:

$$\min_{w} \max_{\theta} \mathbb{E}_{x \sim \mu_{\text{data}}} \left[\log p_\theta(x) \right] + \mathbb{E}_{z \sim \text{Normal}(0, I)} \left[\log(1 - p_\theta(G_w(z))) \right]$$

Putting everything together...



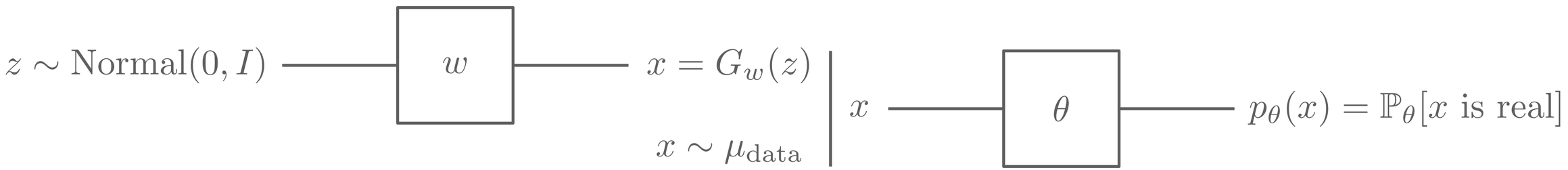
Training a GAN



- We want to solve the minimax problem

$$\min_w \max_\theta \mathbb{E}_{x \sim \mu_{\text{data}}} [\log p_\theta(x)] + \mathbb{E}_{z \sim \text{Normal}(0,I)} [\log(1 - p_\theta(G_w(z)))]$$

Training a GAN



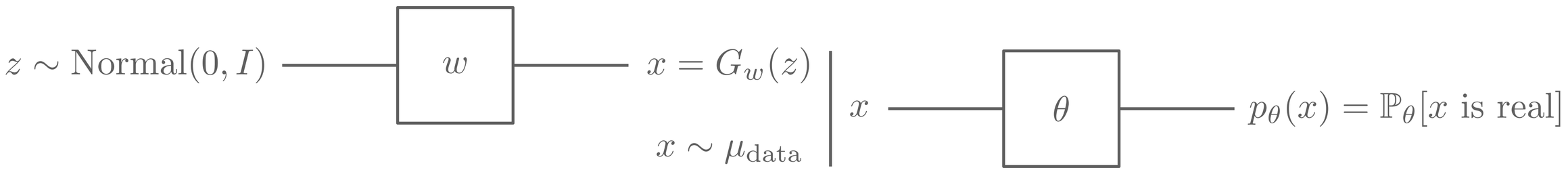
- We use stochastic gradient descent
- We alternate between:
  - Minimizing the generator loss,

$$L_G(w) = \frac{1}{N} \sum_{n=1}^{N} \log(1 - p_\theta(G_w(z_n)))$$

Minibatch size

$z_n \sim \text{Normal}(0, I)$

Training a GAN



- We use stochastic gradient descent
- We alternate between:
  - ... and minimizing the discriminator loss,

$$L_D(\theta) = -\frac{1}{N}\sum_{n=1}^{N} \log p_\theta(x_n) - \log(1 - p_\theta(G_w(z_n)))$$

$z_n \sim \text{Normal}(0, I)$

Sample from dataset

Training a GAN

- Several variants and schedules exist
- In general,
  - The optimization of GANs is often difficult
  - There are no convergence guarantees

Generation using GANs

Generation using GANs

Generation using GANs

Concluding remarks

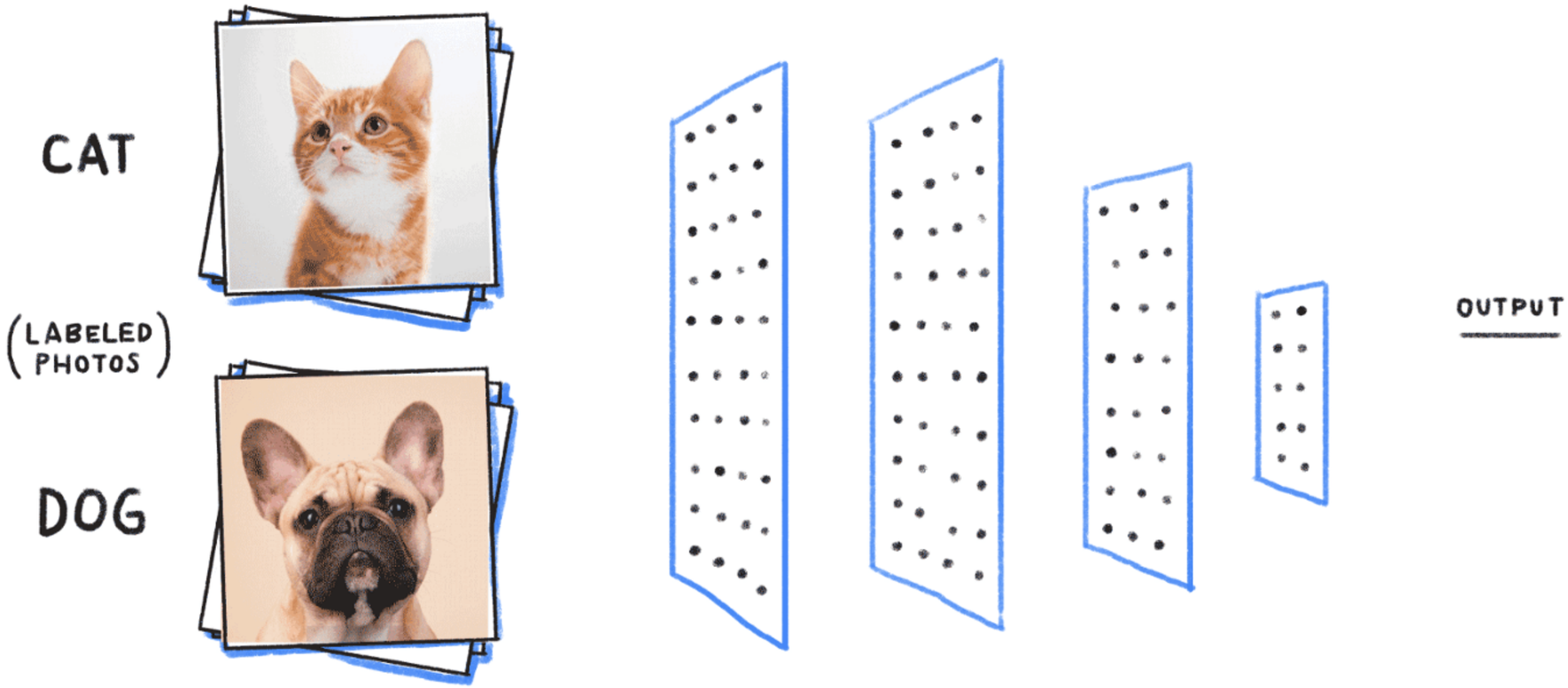
- GANs have several points in their favor:
  - GANs generate the sharpest images
  - They are cheaper to train than other generative models
- However, ...
  - They are difficult to optimize (often unstable training dynamics)
  - We cannot use them to perform statistical inference

The other side of deep learning

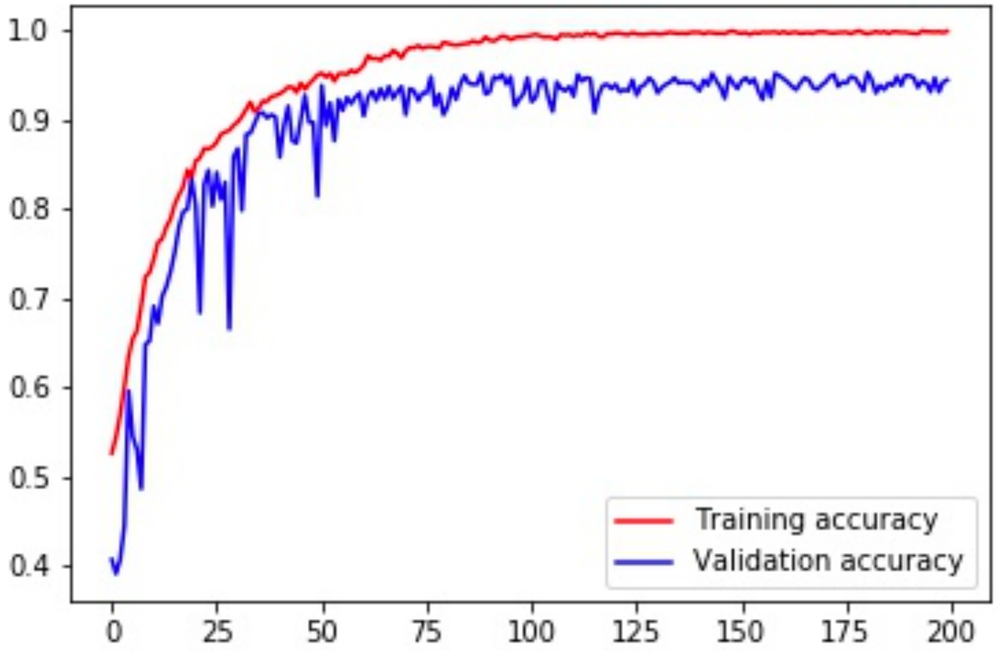
What other side?

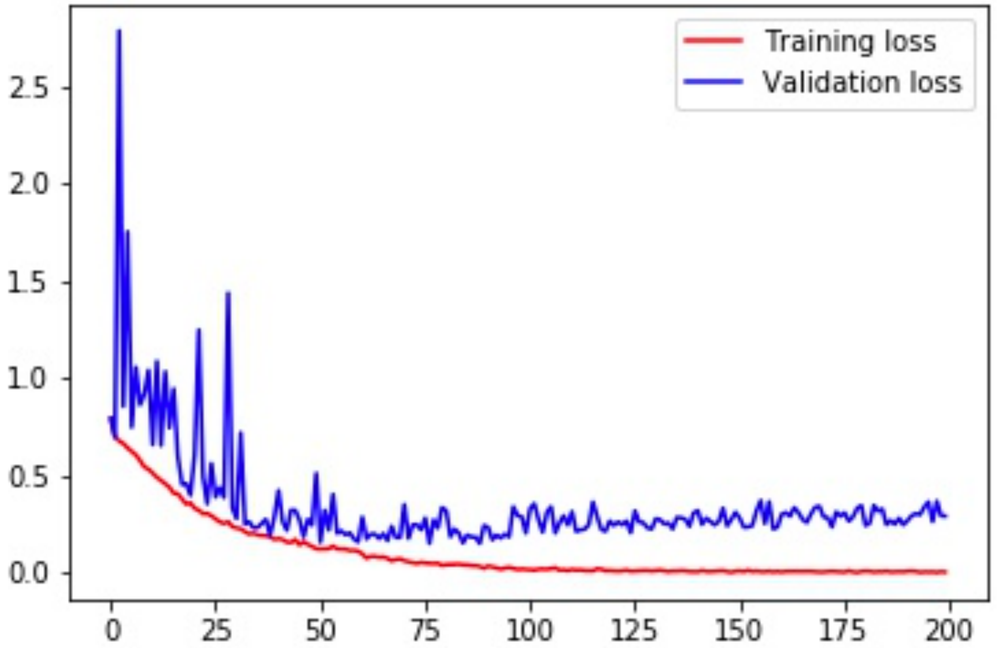
- Are deep learning models trustworthy?
- Can we explain the outputs of a deep learning model?
- Fairness and ethics

Example



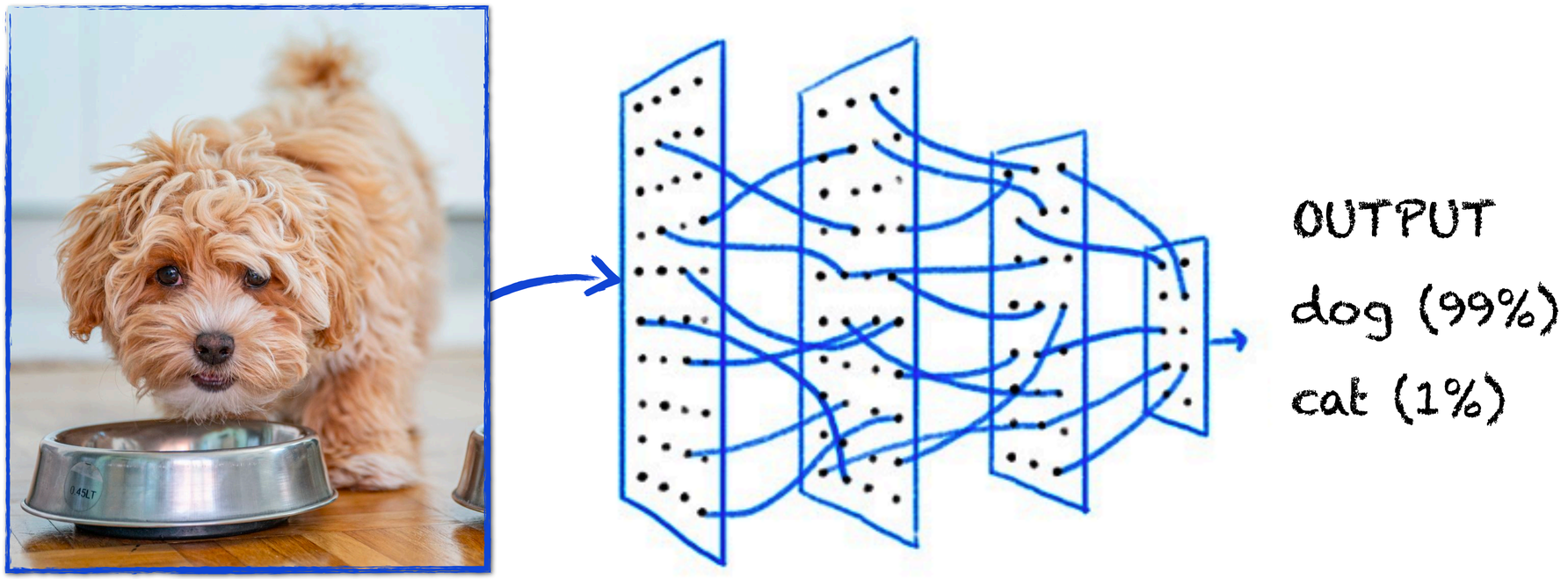
Example


1.0
0.9
0.8
0.7
0.6
0.5
0.4
0
25
50
75
100
125
150
175
200
Training accuracy
Validation accuracy

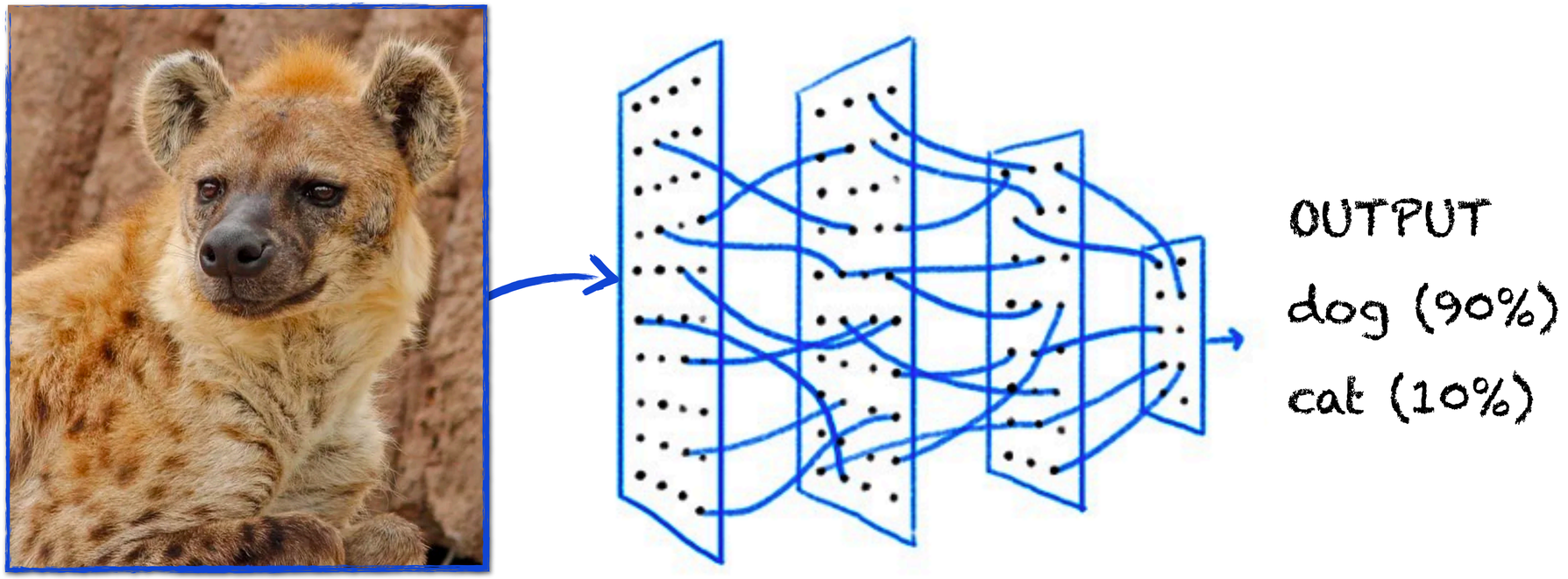

2.5
2.0
1.5
1.0
0.5
0.0
0
25
50
75
100
125
150
175
200
Training loss
Validation loss

Would you trust this model?

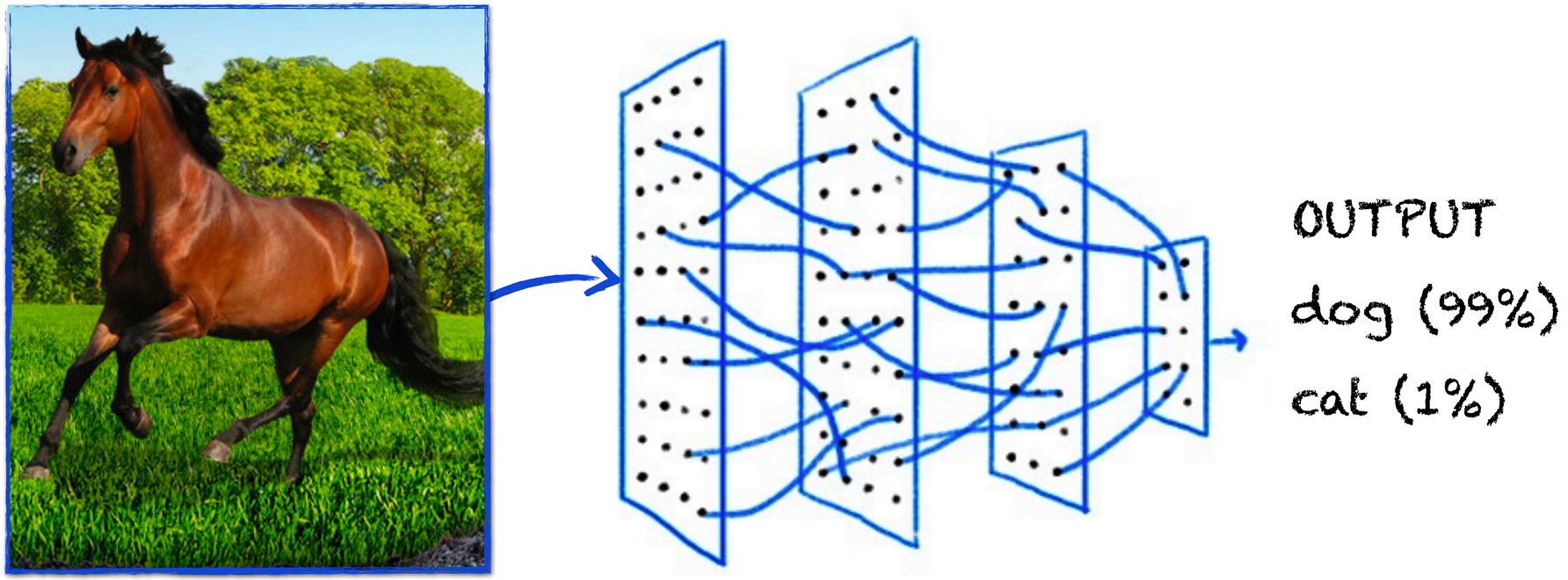
Can we trust deep learning models?



Can we trust deep learning models?



Can we trust deep learning models?



Can we trust deep learning models?

- The model may be expected to be wrong in out-of-domain inputs
  - Can we have some indication about when the model is “confused”?

Uncertainty

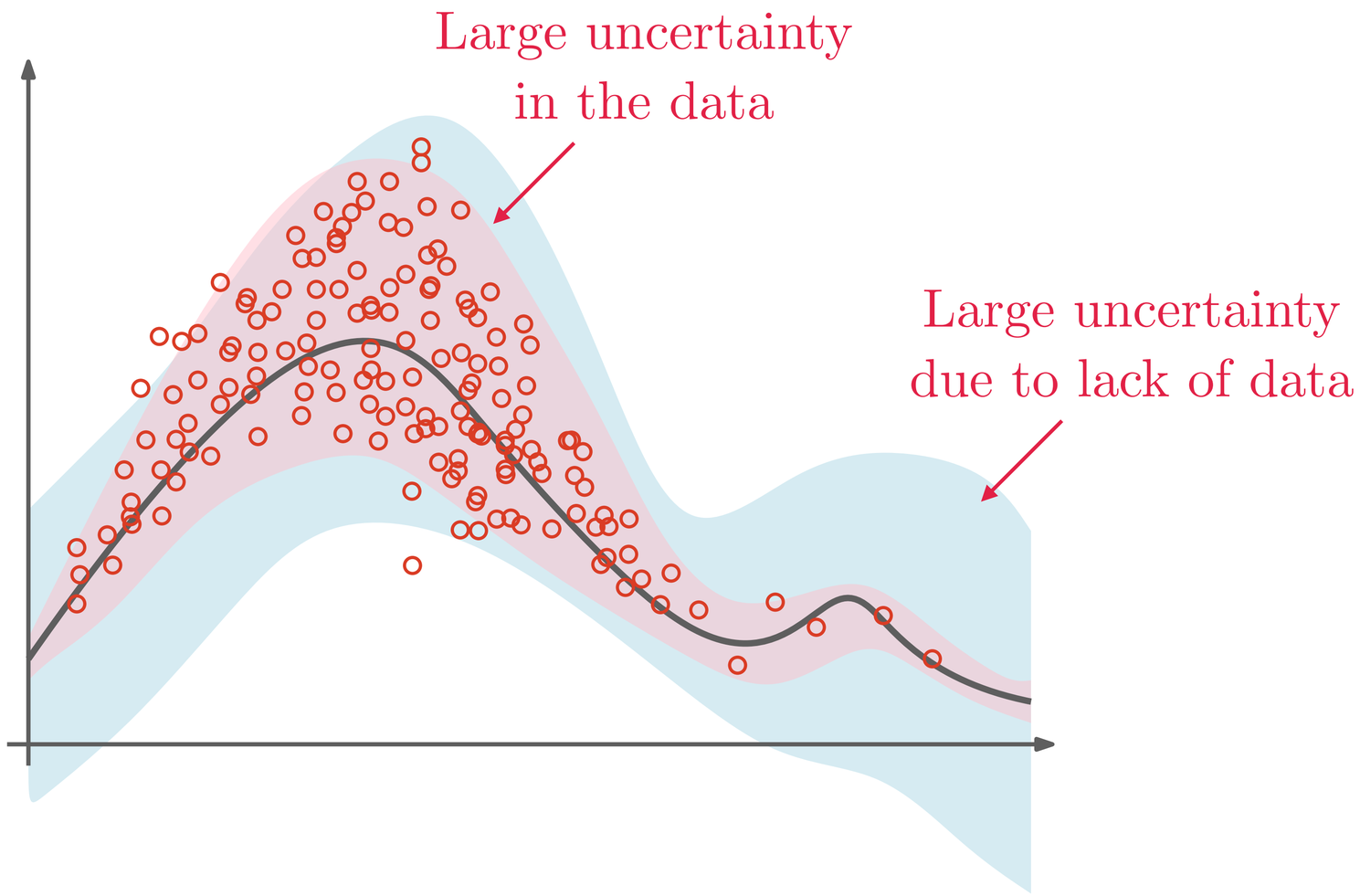
Uncertainty in DL predictions

- Assessing uncertainty in DL predictions is useful for...
  - Deciding when a prediction must be further inspected
  - Debugging, retraining, etc.
  - Providing safe-checks for high-risk applications

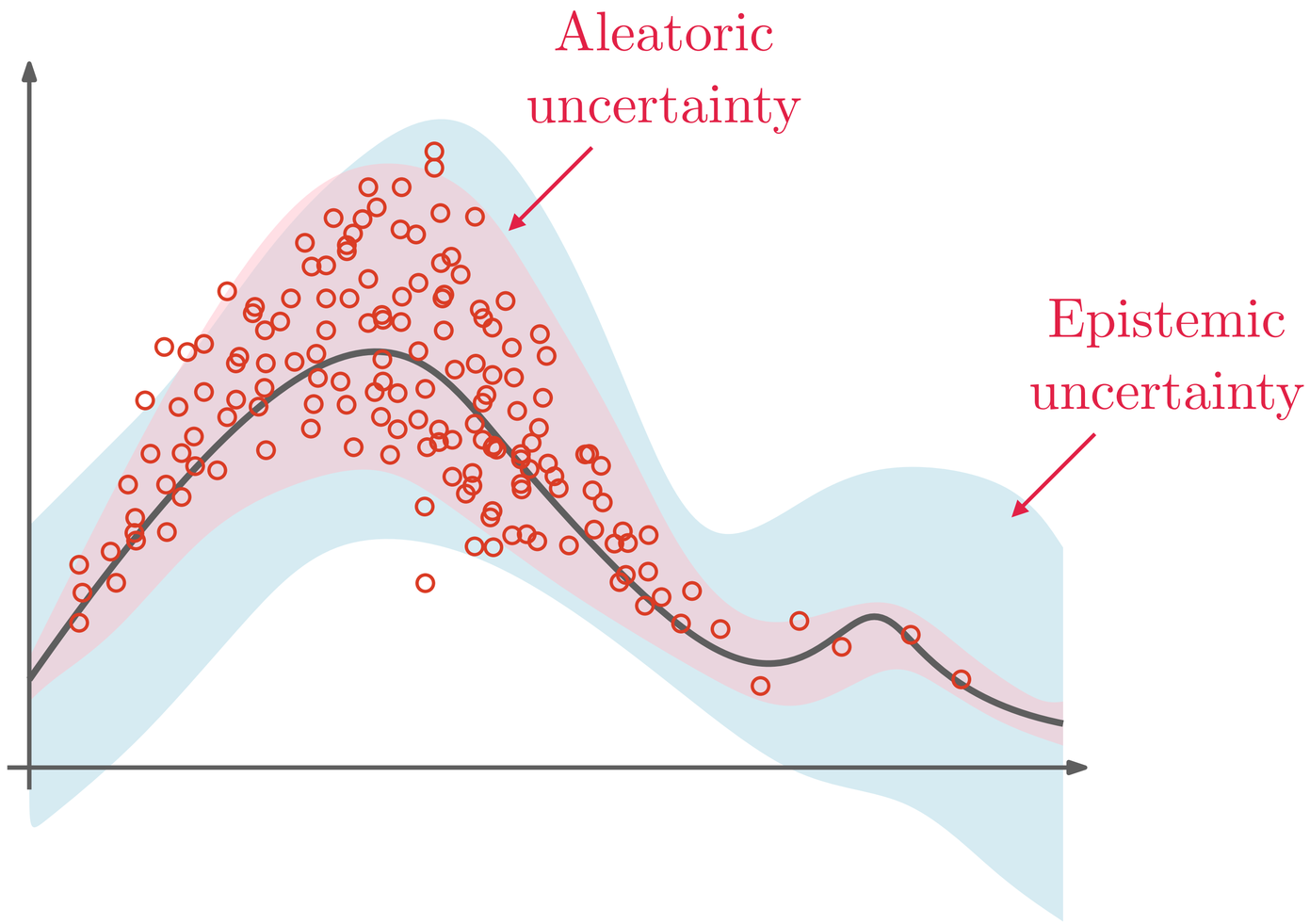
Uncertainty in DL predictions

- Uncertainty in the output of a deep learning model may be caused by:
  - ... uncertainty inherent to the data (aleatoric uncertainty)
  - ... knowledge/model uncertainty (epistemic uncertainty)

Uncertainty in DL predictions



Uncertainty in DL predictions



Uncertainty in DL predictions

Measure of "confusion"

- Estimating aleatoric uncertainty:
  - We can estimate data entropy in the dataset
  - We can use data from multiple annotators to estimate uncertainty

Uncertainty in DL predictions

- Estimating epistemic uncertainty:
  - Often due to lack of data / out-of-domain data
  - How much do predictions change with data?

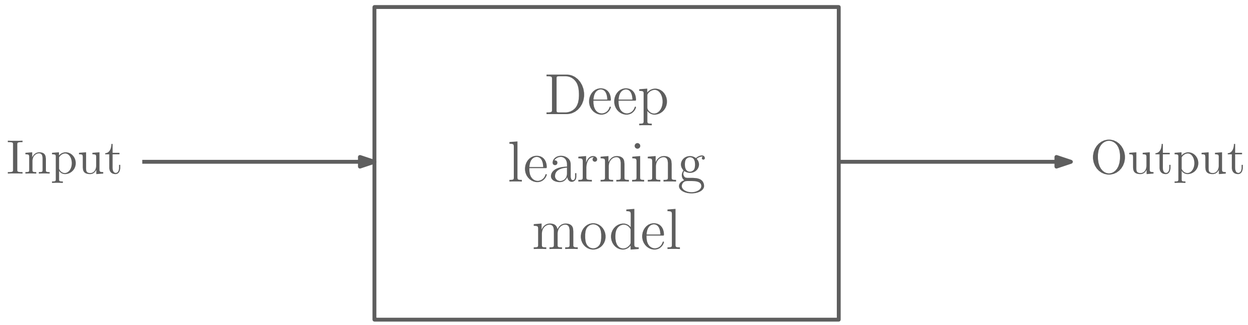
Variance

Variance estimation in DL models

- Estimating variance:
  - Monte-Carlo Dropout (use dropout at estimation time)
  - Deep ensembles (independently training multiple models)
  - ...

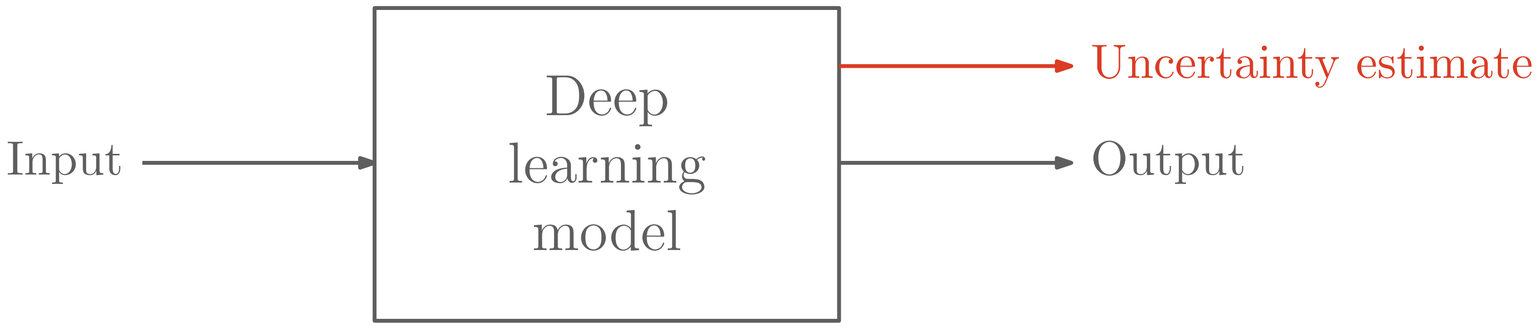
Beyond uncertainty

- Uncertainty provides a way of quantifying to what extent predictions can be trusted



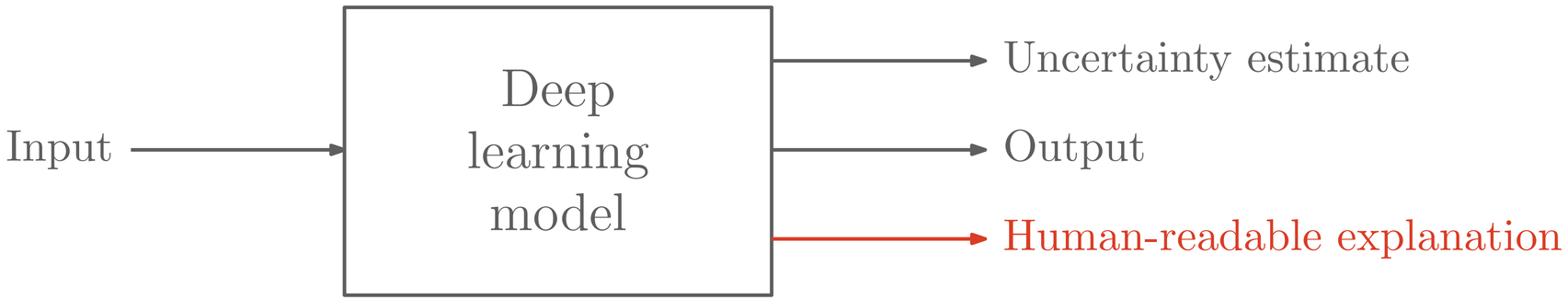
Beyond uncertainty

- Uncertainty provides a way of quantifying to what extent predictions can be trusted



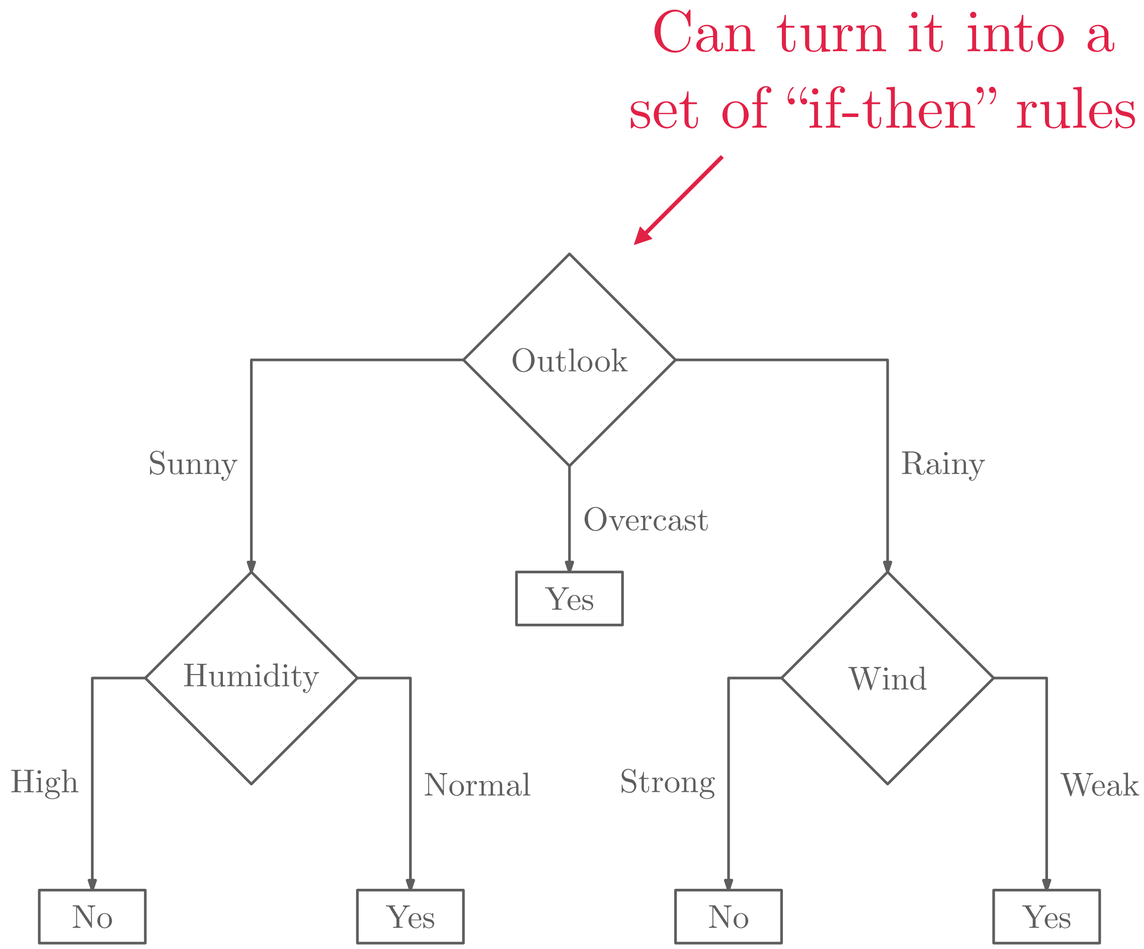
Beyond uncertainty

- Uncertainty provides a way of quantifying to what extent predictions can be trusted
- However, it is often not enough to understand (and trust) the predictions made by DL models



Explainability in machine learning

- Some machine learning models are intrinsically explainable
- E.g.,
  - Decision trees



Explainability in machine learning

- Some machine learning models are intrinsically explainable
- E.g.,
  - Decision trees

If the outlook is overcast, then yes.

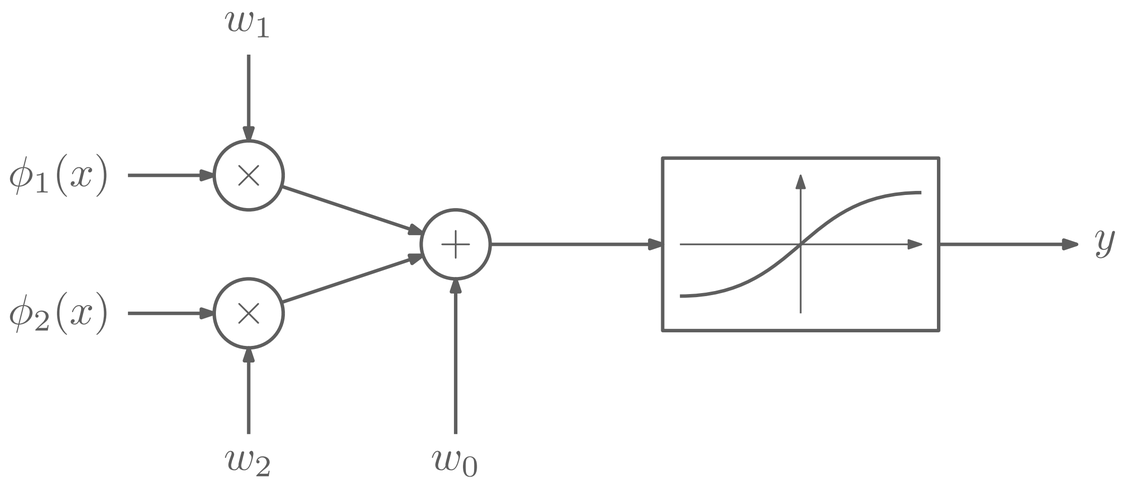
Similarly, if it's sunny but the humidity is normal, then yes.

Same thing if it's rainy but the wind is weak, then yes.

Otherwise, no.

Explainability in machine learning

- Some machine learning models are intrinsically explainable
- E.g.,
  - Decision trees
  - Linear models



$$y = F(w_0 + w_1\phi_1(x) + w_2\phi_2(x))$$

Relative importance of different features

Can we have post-hoc explanations?

- We can "build" local explanations for a trained model

Explanations computed
for specific inputs

Can we have post-hoc explanations?

- We can "build" local explanations for a trained model
  - Model-specific explanations (can be applied to specific models)
  - Model-agnostic explanations (can be applied to "arbitrary" models)

Model-specific explanations

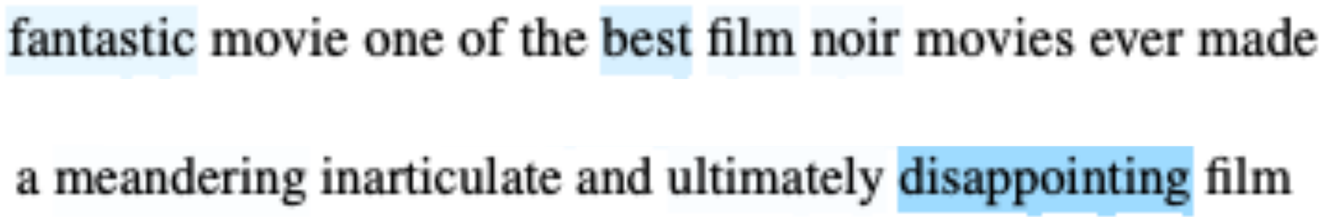
- **Gradient-based explanations**
  - Can be used in differentiable machine learning models (e.g., neural networks)
  - E.g., use gradients to reveal which regions of input contribute the most to the prediction

$$F_w(x+\epsilon) \approx F(x) + \epsilon^\top \nabla_x F_w(x) + \text{small stuff}$$

Gradient w.r.t. input indicates
which input directions (features)
most impact the output

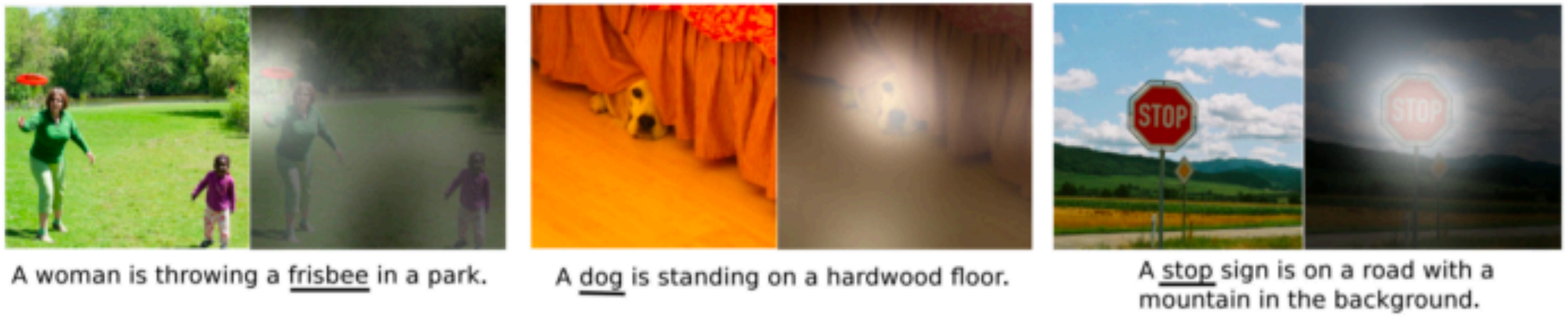
Model-specific explanations

- **Attention-based explanations**
  - Can be used with attention-based models
  - E.g., use attention weights to determine most relevant inputs



Model-specific explanations

- **Attention-based explanations**
  - Can be used with attention-based models
  - E.g., use attention weights to determine most relevant inputs



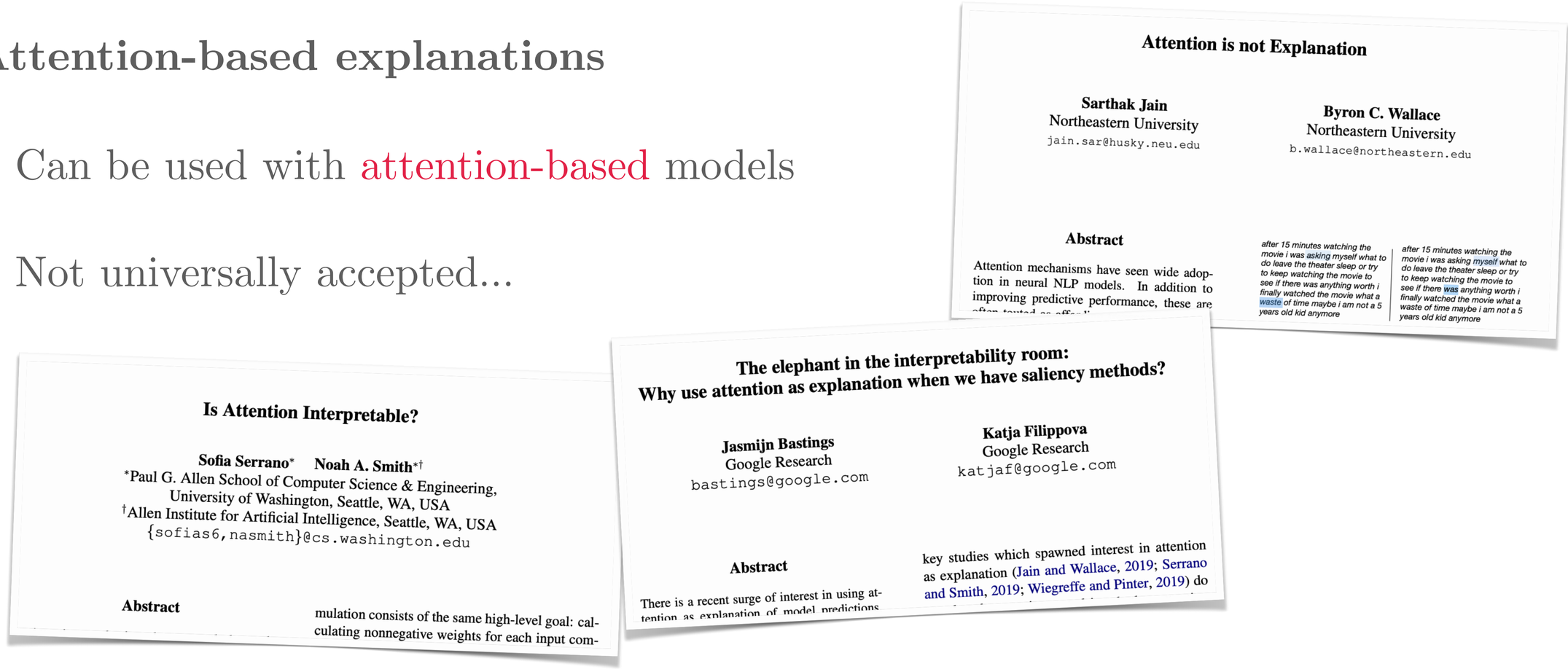
A woman is throwing a frisbee in a park.

A dog is standing on a hardwood floor.

A stop sign is on a road with a mountain in the background.

Model-specific explanations

- **Attention-based explanations**
  - Can be used with attention-based models
  - Not universally accepted...



Attention is not Explanation

Sarthak Jain
Northeastern University
jain.sar@husky.neu.edu

Byron C. Wallace
Northeastern University
b.wallace@northeastern.edu

Abstract

Attention mechanisms have seen wide adoption in neural NLP models. In addition to improving predictive performance, these are

after 15 minutes watching the movie i was asking myself what to do leave the theater sleep or try to keep watching the movie to see if there was anything worth i finally watched the movie what a waste of time maybe i am not a 5 years old kid anymore

after 15 minutes watching the movie i was asking myself what to do leave the theater sleep or try to keep watching the movie to see if there was anything worth i finally watched the movie what a waste of time maybe i am not a 5 years old kid anymore

The elephant in the interpretability room:
Why use attention as explanation when we have saliency methods?

Jasmijn Bastings
Google Research
bastings@google.com

Katja Filippova
Google Research
katjaf@google.com

Abstract

There is a recent surge of interest in using attention as explanation of model predictions

key studies which spawned interest in attention as explanation (Jain and Wallace, 2019; Serrano and Smith, 2019; Wiegreffe and Pinter, 2019) do

Is Attention Interpretable?

Sofia Serrano* **Noah A. Smith*†**
*Paul G. Allen School of Computer Science & Engineering,
University of Washington, Seattle, WA, USA
†Allen Institute for Artificial Intelligence, Seattle, WA, USA
{sofias6,nasmith}@cs.washington.edu

Abstract

mulation consists of the same high-level goal: calculating nonnegative weights for each input com-

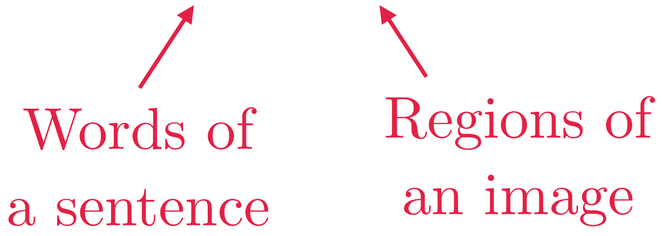
Model-agnostic explanations

- **Perturbation-based explanations**
  - Share intuition with gradient-based explanations
  - Seek to answer the question:

    "Which part of the input most influences the prediction?"
  - E.g., LIME (Local Intepretable Model-agnostic Explanations)

LIME

- For given input x,
  - Select features of the input



LIME

- For given input x,
  - Select features of the input
  - Generate neighbors of x by "perturbing" features

Mask/hide words of a sentence

Add noise to pixels in a region

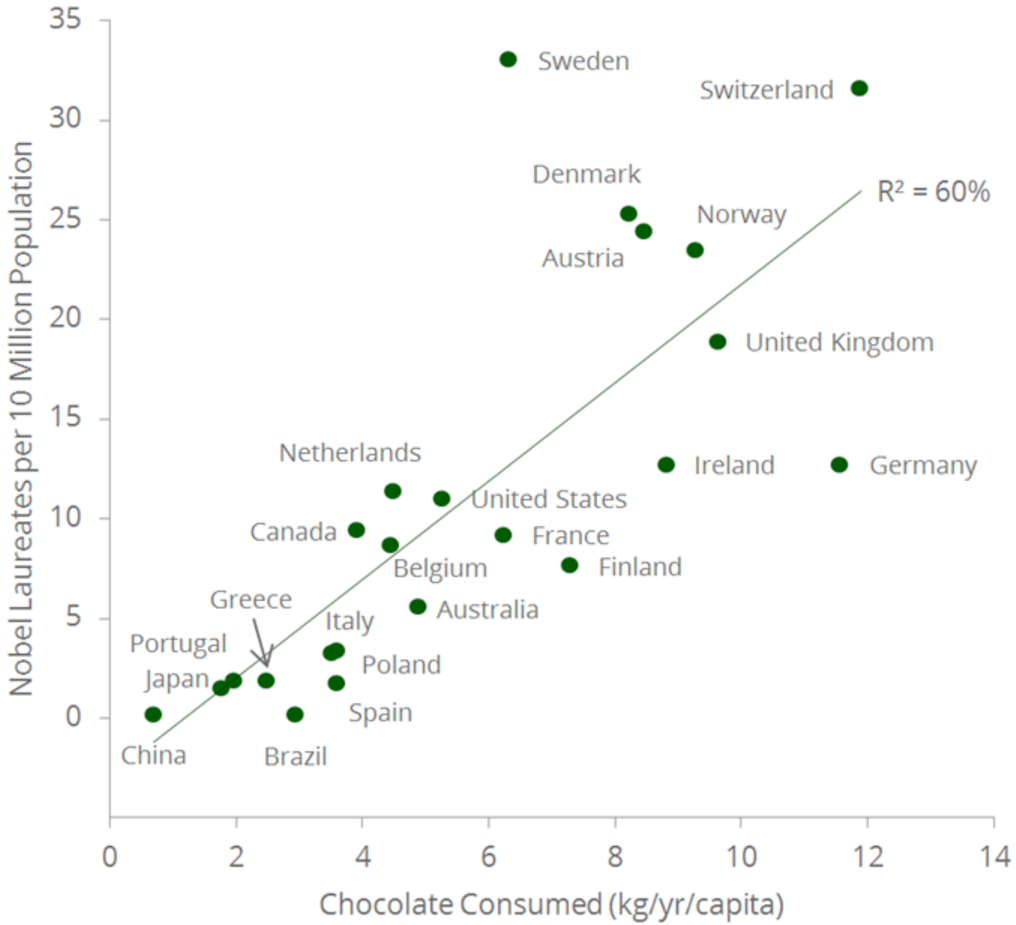
LIME

- For given input x,
  - Select features of the input
  - Generate neighbors of x by "perturbing" features
  - Train "interpretable model" (e.g., decision tree, linear model) to mimic the output of the original model in the generated neighbors

Correlation

Correlation is not causation

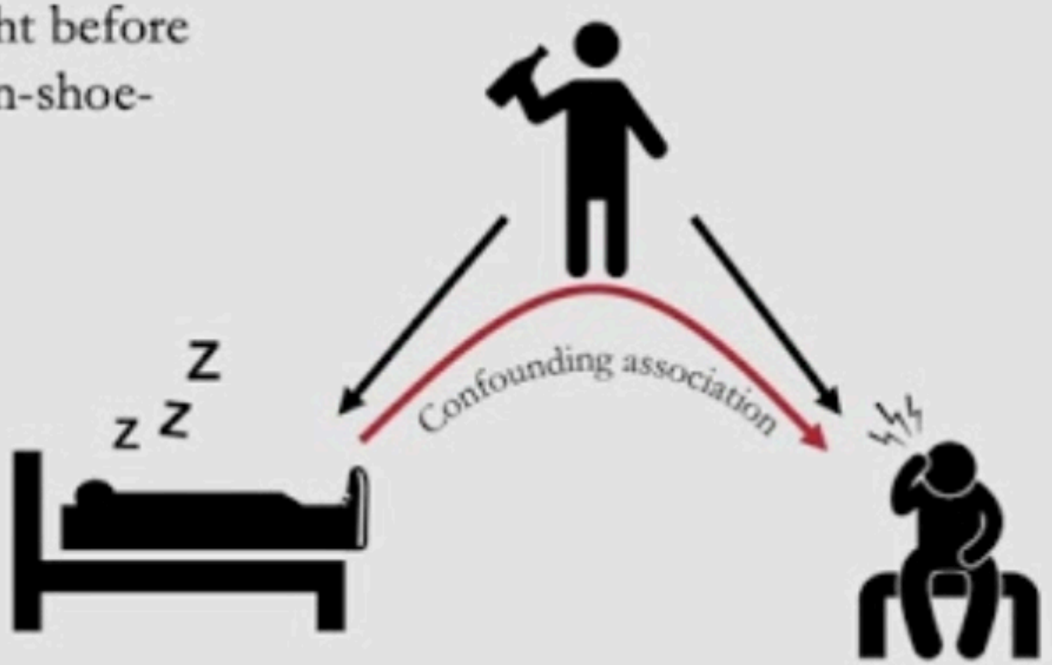
99% of models used (linear or non-linear) make decision based on correlations

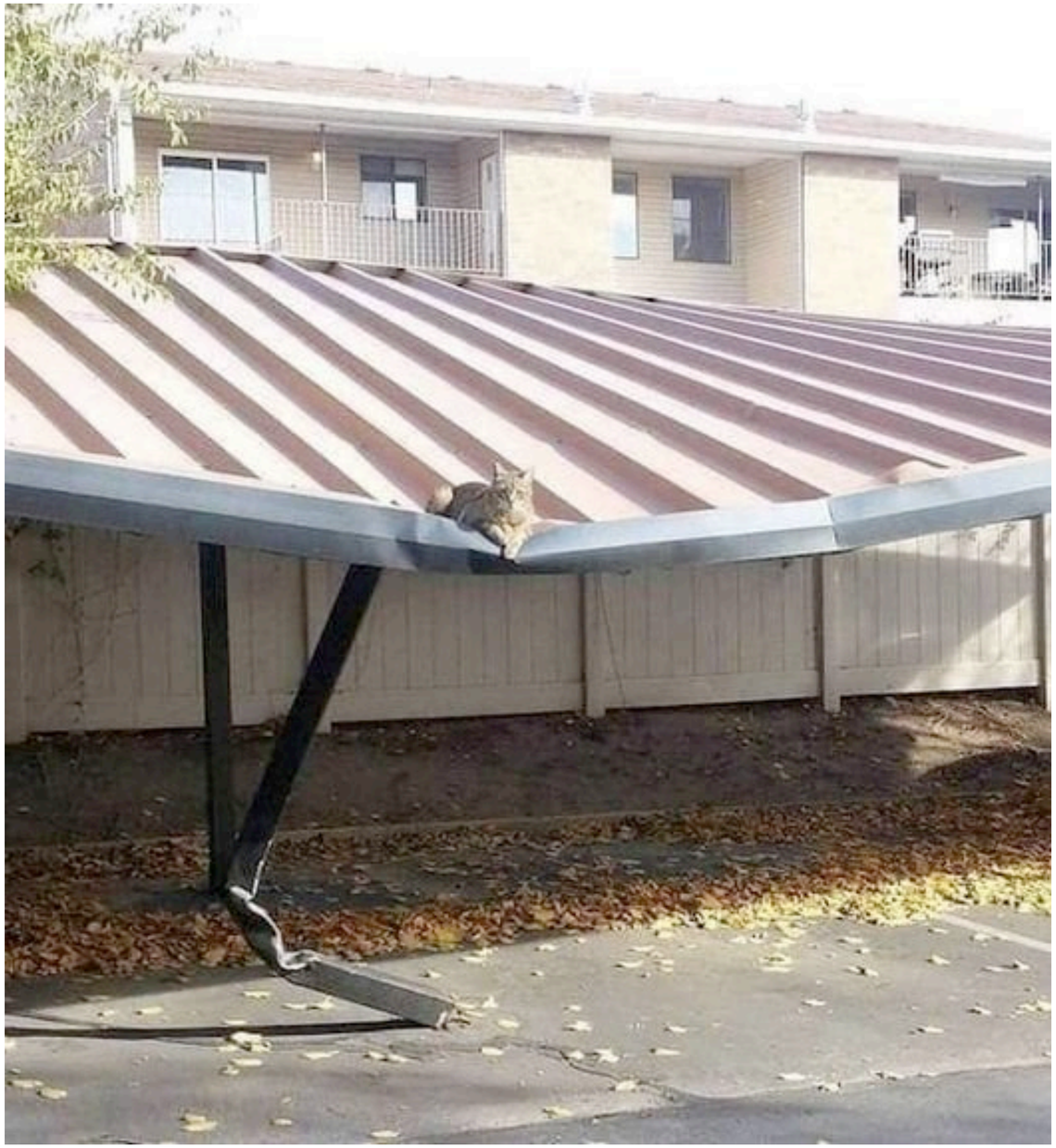


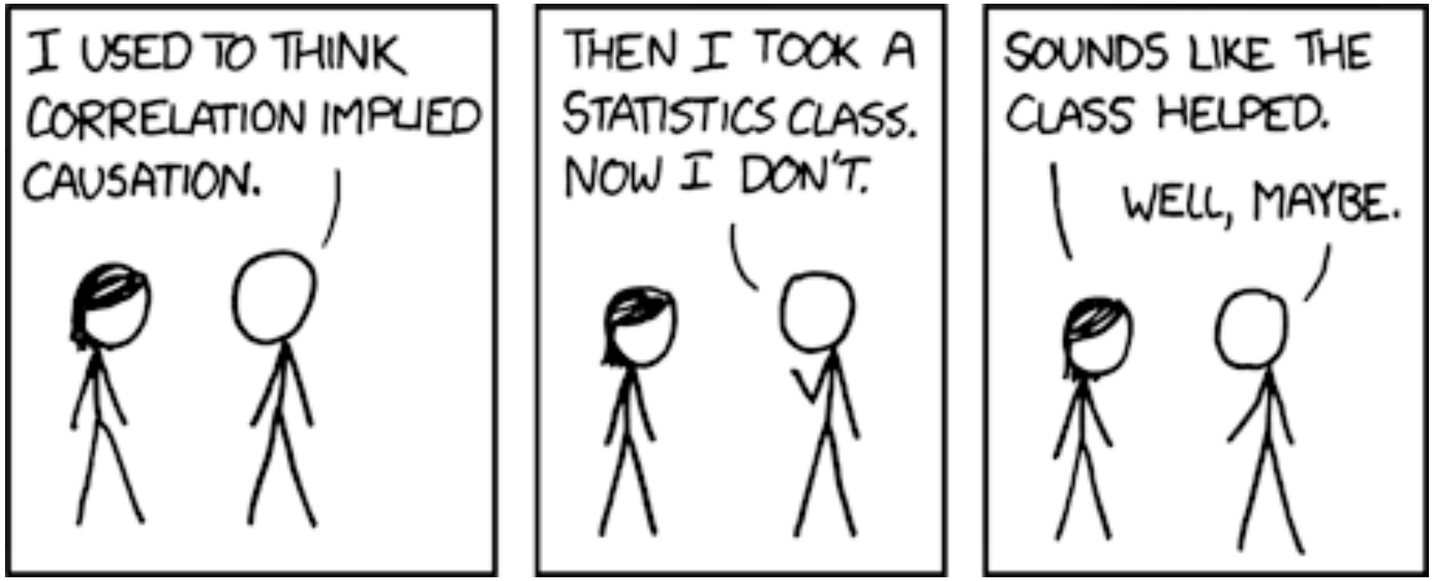
Sleeping with shoes on is strongly correlated with waking up with a headache

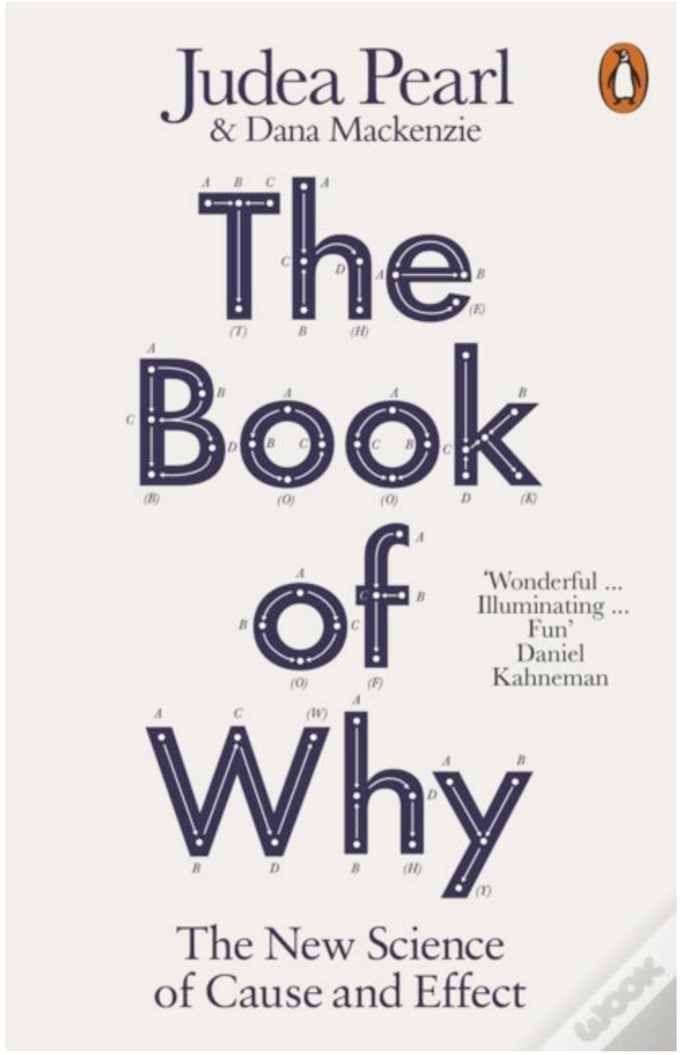
Common cause: drinking the night before

1. Shoe-sleepers differ from non-shoe-sleepers in a key way
2. Confounding





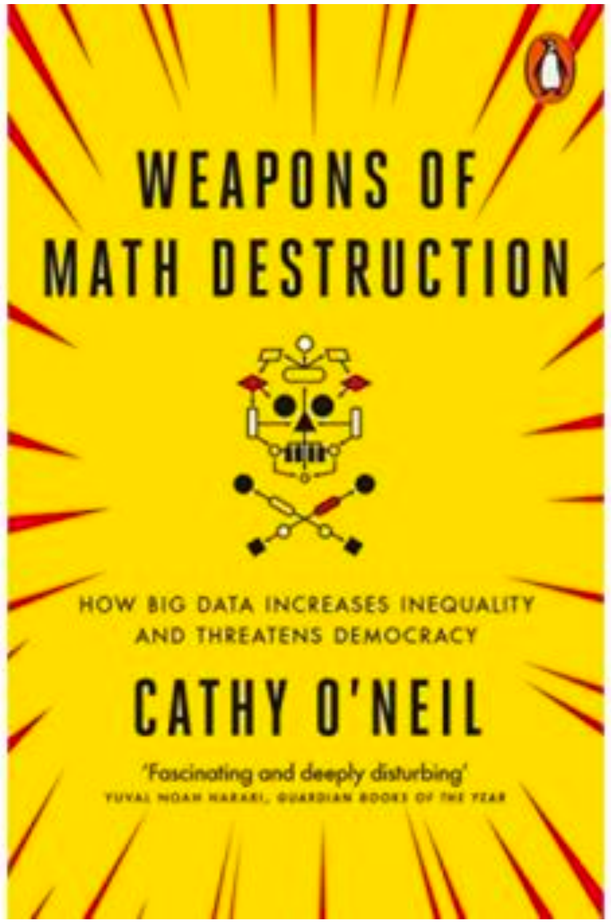

I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

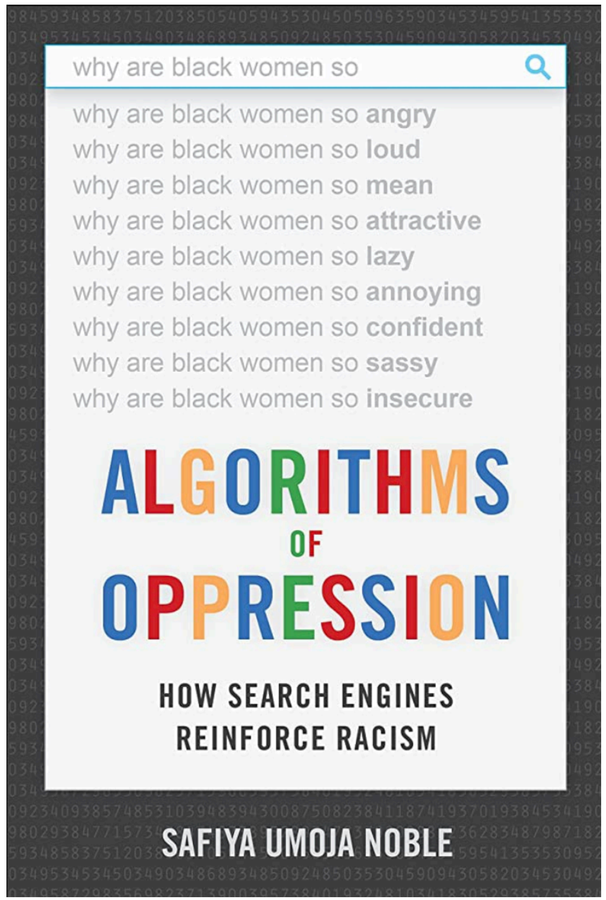

Judea Pearl
& Dana Mackenzie
The Book of Why
'Wonderful ... Illuminating ... Fun'
Daniel Kahneman
The New Science of Cause and Effect

Ethics and fairness

Ethics

Beyond performance, trust, and explainability





Ethics

Beyond performance, trust, and explainability

- The extensive use of deep learning models in different aspects of our daily lives raises important ethical concerns
  - Discrimination bias
  - Accessibility
  - Privacy compromise
  - Sustainability

Bias

- What is bias?
  - Statistical bias is a systematic tendency which induces differences between results and facts
  - Inductive bias (in machine learning) is the set of assumptions that an algorithm uses to make predictions about inputs it has never encountered

Bias

- What is bias?
  - Representation bias (aka sample bias) is when the training dataset does not include a balanced representation of instances appearing in the test set.
  - Confounding bias is when there is a distortion of the relation between independent and dependent variables due to a third variable that is independently related to both

Bias

- What is bias?
  - Association bias is when spurious implicit associations occur in the training data but not in the test data
  - Observer bias (aka confirmation bias) is when subjective judgements influence the process of data-handling (we "see only what we expect to see")
  - Exclusion bias is when data is removed for being "unimportant"

Bias

- What is bias?
  - **Discrimination bias** (aka demographic bias or racial bias) is when the data is skewed to favor a particular demographic group

Discrimination bias

CORPORATE TALENT & INCLUSION

New study finds AI-enabled anti-Black bias in recruiting

Dawn Zapata Senior Content Producer / Thomson Reuters

18 Jun 2021

Amazon scraps secret AI recruiting tool that showed bias against women

Jeffrey Dastin

8 MIN READ

NOVEMBER 1, 2018

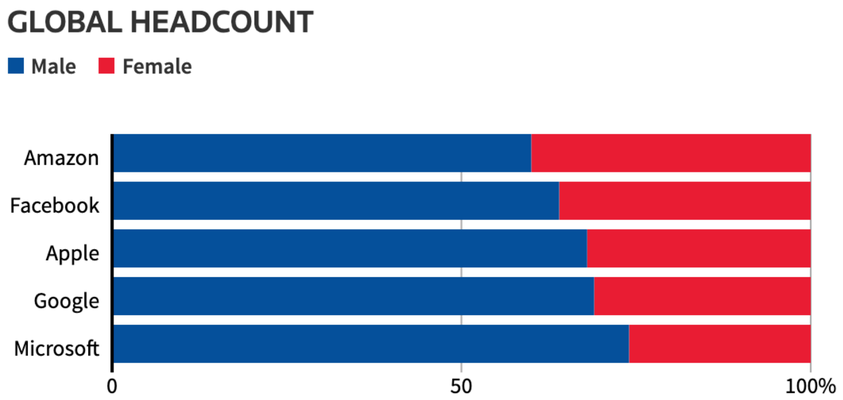
Amazon's sexist hiring algorithm could still be better than a human

by Maude Lavanchy, The Conversation

MPW • AMAZON

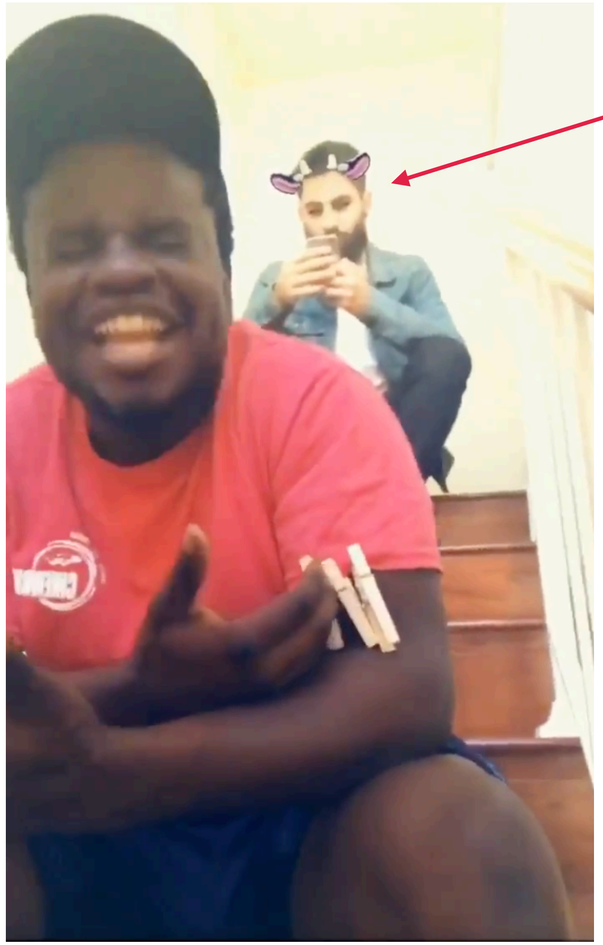
Amazon Reportedly Killed an AI Recruitment System Because It Couldn't Stop the Tool from Discriminating Against Women

By David Meyer October 10, 2018



Discrimination bias

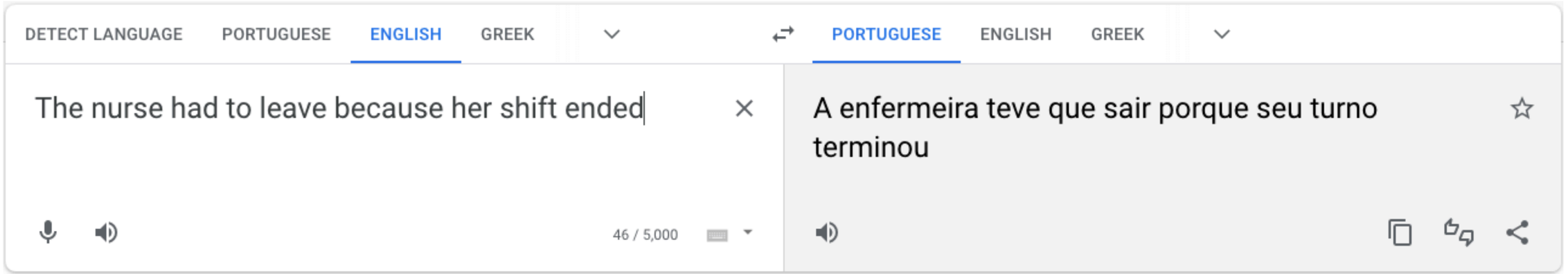
Discrimination bias

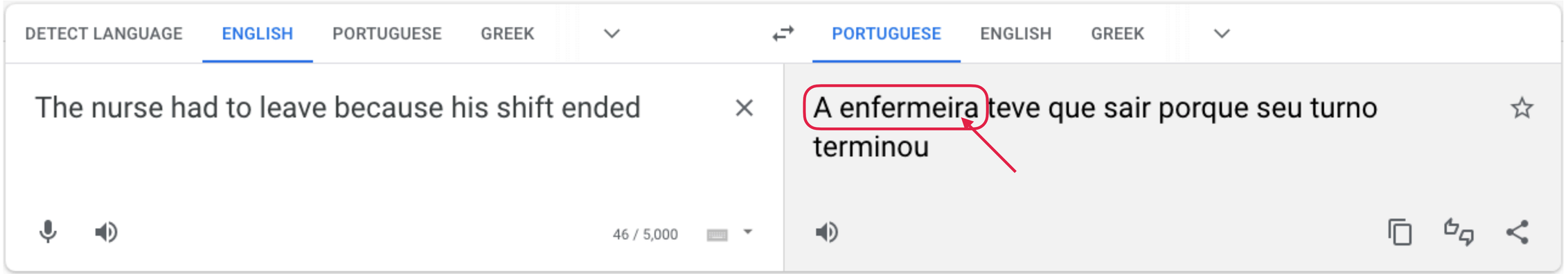


Snapchat filters don't work on dark skin

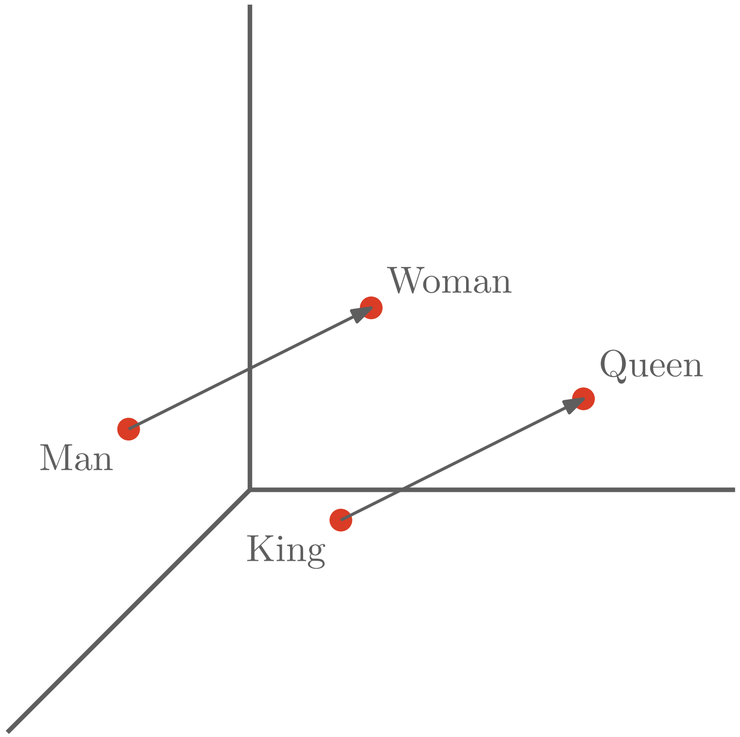
Soap dispensers don't work on dark skin

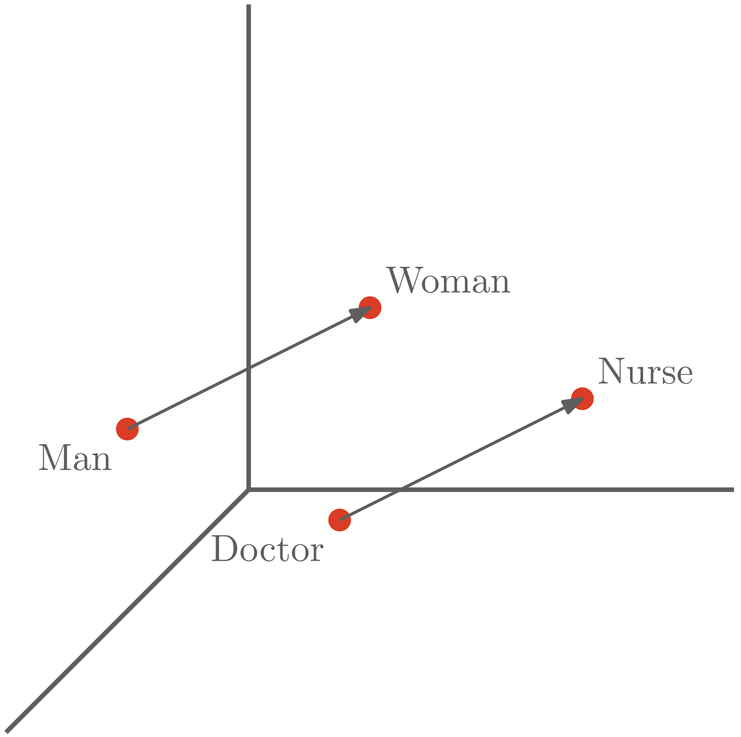
Discrimination bias



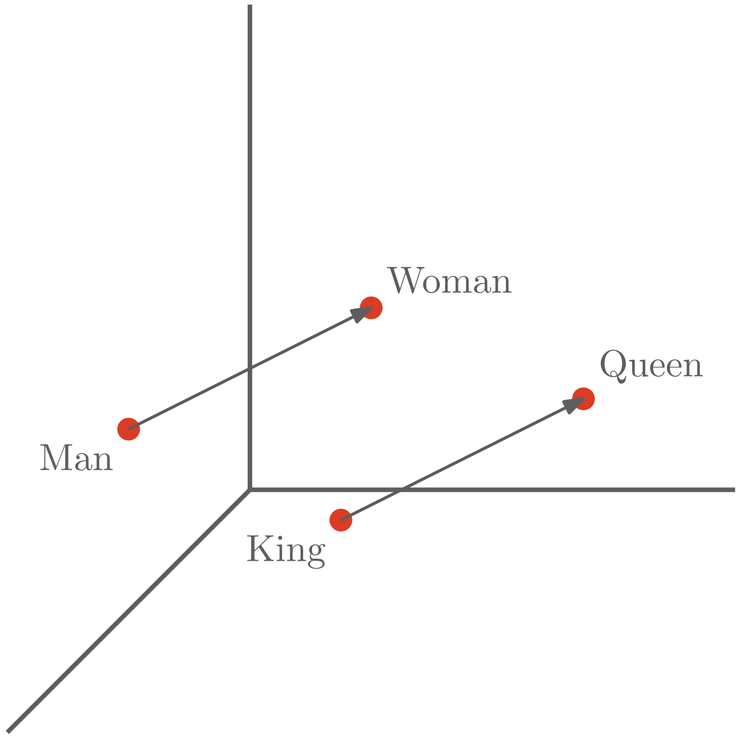


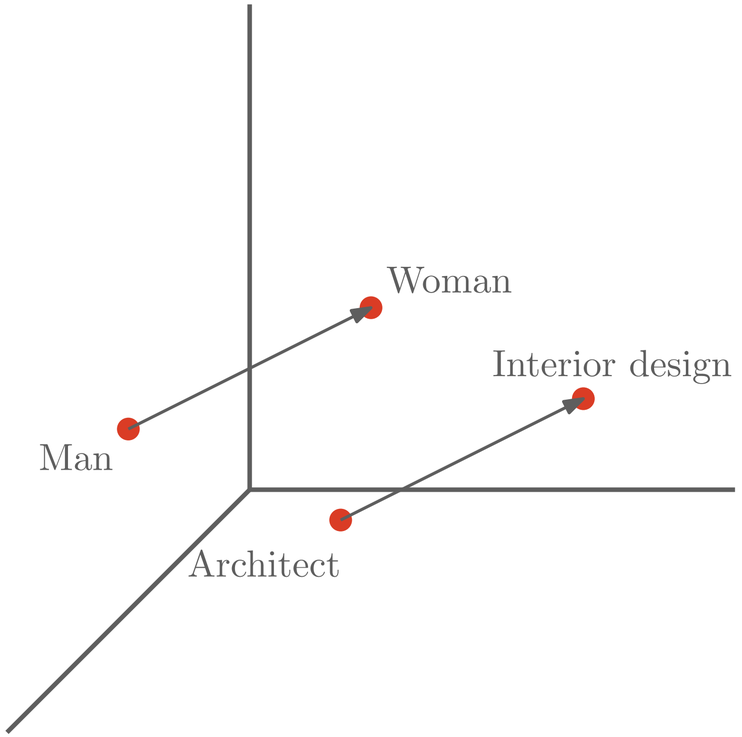
Biased language models





Biased language models





How do we correct biases?

Pre-training:

- Use diverse datasets
- Use social context during annotation
- ...

How do we correct biases?

Training:

Assume feature vectors can be decomposed as

Gender information

$$\phi(x) = \phi_{\perp}(x) + \phi_g(x)$$

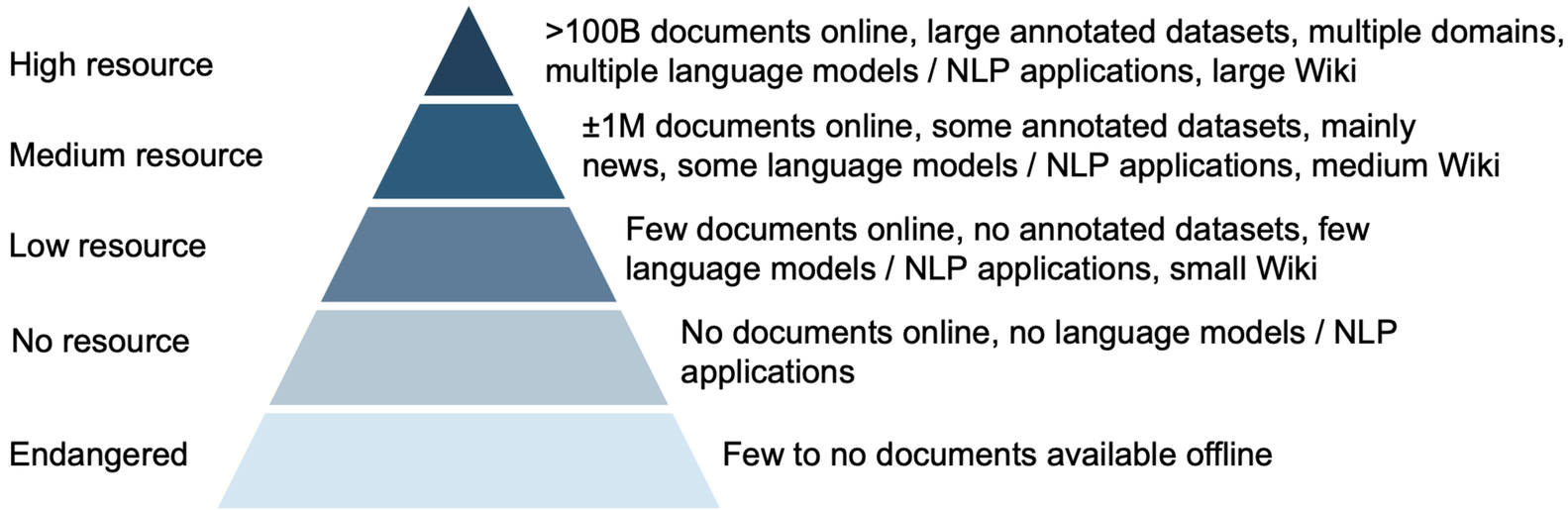
- Loss terms that restrict use of gender information
- Regularization terms that penalize projection in gender dimensions

Accessibility

- Who has access to ML-driven applications?
- Who are ML-driven applications designed for?

An example: NLP

- Are NLP models available for all languages?
- Are multilingual models (e.g., machine translation models) equally efficient/effective for all languages?



Privacy

When is it a concern?

- Should we be tracked?
- Should we be tracked if we consent?
- Even if we consent, how informed is the consent? Do we know how data will be used?
- What if we change our mind?

Privacy

Mumbai-based Agency Leaked Data Of Over 49 Million Instagram Influencers And

The New York Times

Cambridge Analytica and Facebook: The Scandal and the Fallout So Far

Revelations that digital consultants to the Trump campaign misused the data of millions of Facebook users set off a furor on both sides of the Atlantic. This is how The Times covered it.

CPO MAGAZINE

HOME NEWS INSIGHTS RESOURCES

DATA PRIVACY NEWS · 4 MIN READ

Big Tech Isn't Breaking Any Privacy Rules if There Aren't Rules to Break

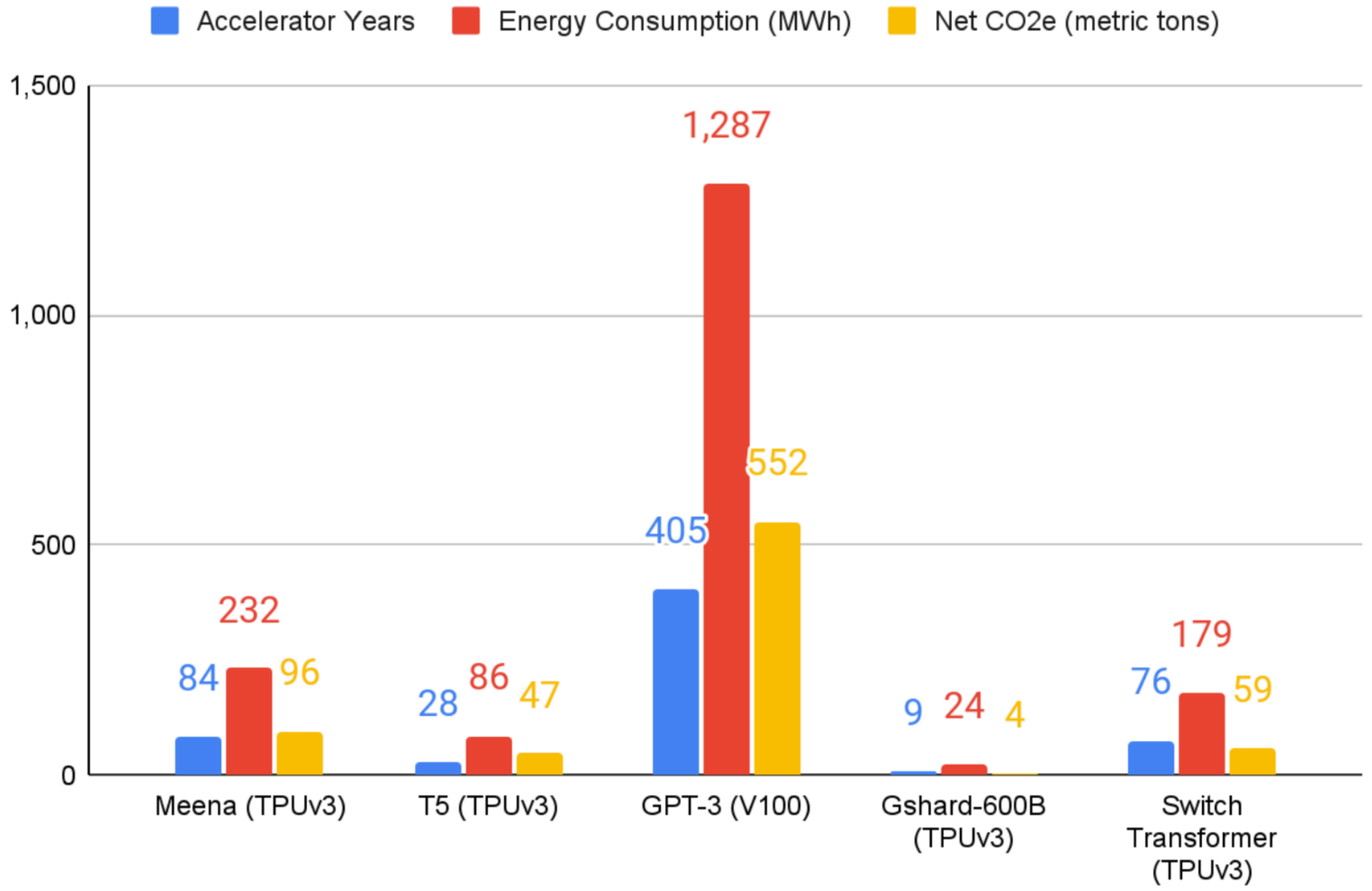
DIMITRI SHELEST · DECEMBER 27, 2021

Sustainability

- Big AI and deep learning models consume considerable resources
  - Human resources (engineering, data curation, annotation, etc.)
  - Material resources (hardware)
  - Power consumption
  - CO2 emissions

Training BERT:
Trans-american jet flight

Sustainability



Sustainable AI

- Consider sustainability at every step of the AI/DL life cycle
  - Keep humans in the loop
- AI can contribute to the development of sustainable solutions towards several of the UN Sustainable Development Goals

Summary

- Generative models are useful to model high-dimensional data
- Both VAEs and GANs fall into the family of differentiable generator networks
- VAEs are trained by maximizing the ELBO
- GANs are trained adversarially

Summary

- The generalization of deep learning technology poses several challenges:
  - How can we develop trustworthy/interpretable models?
  - How can we ensure that deep learning technology development is...
    - ... unbiased?
    - ... accessible?
    - ... sustainable?

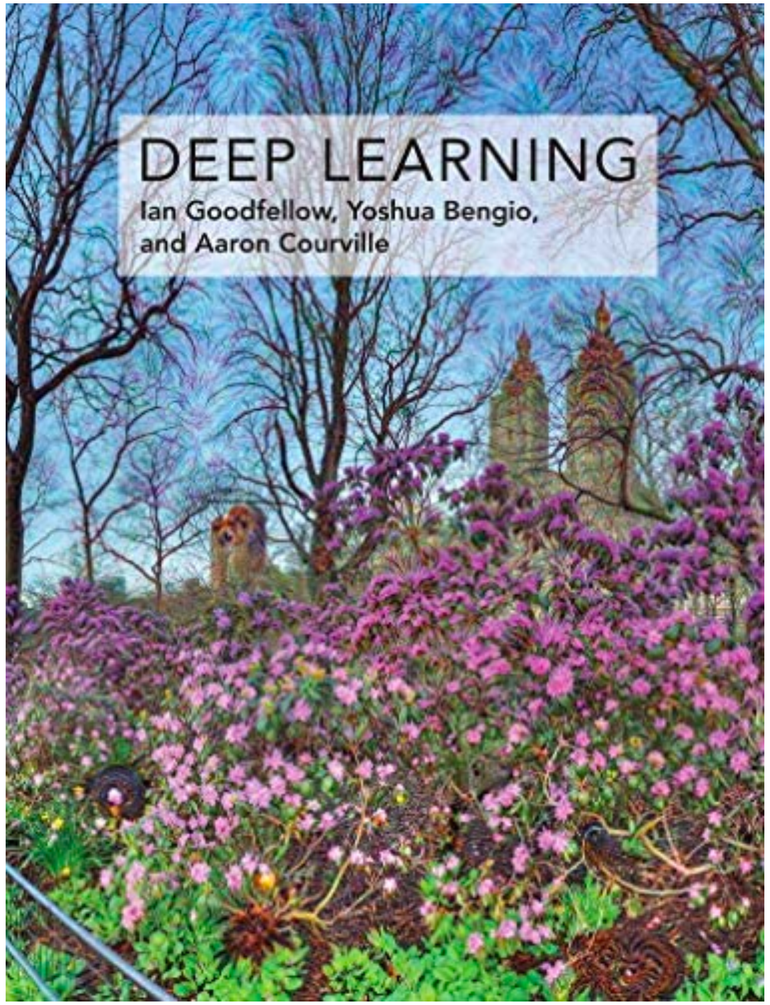
To conclude...

- In the course we covered:
  - Feed-forward neural networks
  - Convolutional neural networks
  - Recurrent neural networks
  - Sequence-to-sequence models and attention mechanisms
  - Generative models

Acknowledgements

A lot of the material used in the course is based on the slides from the MSc course on Deep Learning at Instituto Superior Técnico, authored by André Martins, Mário Figueiredo, and Chrysoula Zerva

References


DEEP LEARNING
Ian Goodfellow, Yoshua Bengio,
and Aaron Courville

References

- Bach, S., Binder, A., Montavon, G., Klauschen, F., Muller, K., and Samek, W. "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation." *PloS One*, 10(7):e0130140, 2015.
- Bordia, S. and Bowman, S. "Identifying and reducing gender bias in word-level language models." *CoRR*, abs/1904.03035, 2019.
- Caliskan, A., Bryson, J., and Narayanan, A. "Semantics derived automatically from language corpora contain human-like biases." *Science*, 356(6334):183-186, 2017.
- Gal, Y., and Z. Ghahramani. "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning." In *Proc. 33rd Int. Conf. Machine Learning*, pp. 1050-1059, 2016.
- Gonen, H. and Goldberg, Y. "Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them." *CoRR*, abs/1903.03862, 2019.

References

- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. "Generative adversarial nets." In *Adv. Neural Information Processing Systems 27*, pp. 2672-2680, 2014.
- Kingma, D., and Welling, M. "Auto-encoding variational Bayes." *CoRR*, abs/1312.6114, 2013.
- Lakshminarayanan, B., Pritzel, A., and Blundell, C. "Simple and scalable predictive uncertainty estimation using deep ensembles." In *Adv. Neural Information Processing Systems 30*, pp. 6405-6416, 2017.
- Ribeiro, M., Singh, S., and Guestrin, C. "Why should I trust you? Explaining the predictions of any classifier." In *Proc. 22nd ACM-SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, pp. 1135-1144, 2016.
- Selvaraju, R., Das, A., Vedantam, R., Cogswell, M., Parikh, D., and Batra, D. "Grad-CAM: Why did you say that?" *CoRR*, abs/1611.07450, 2016.
- Vashishth, S., Upadhyay, S., Tomar, G., and Faruqui, M. "Attention interpretability across NLP tasks." *CoRR*, abs/1909.11218, 2019.